Genome Biology

RESEARCH Open Access

# Methylome-wide association studies and epigenetic biomarker development for 133 mass spectrometry-assessed circulating proteins in 14,671 Generation Scotland participants

Josephine A. Robertson[1], Jakub Bajzik[2], Spyros Vernardis[3,4], Aleksandra D. Chybowska[1], Daniel L. McCartney[1], Arturas Grauslys[4], Jure Mur[1,8], Hannah M. Smith[1], Archie Campbell[1,12], Camilla Drake[5], Hannah Grant[1], Jamie Pearce[6], Tom C. Russ[7,8], Poppy Adkin[3,9], Matthew White[3], Charles Brigden[4], Christoph B. Messner[3,13], David J. Porteous[1], Caroline Hayward[1,5], Simon R. Cox[10], Aleksej Zelezniak[3,4,14,15,16], Markus Ralser[3,4,11], Matthew R. Robinson[2] and Riccardo E. Marioni[1*]

*Correspondence:
Riccardo E. Marioni
riccardo.marioni@ed.ac.uk

Full list of author information is available at the end of the article

## Abstract

**Background** DNA methylation (DNAm) can regulate gene expression, and its genome-wide patterns (epigenetic scores or EpiScores) can act as biomarkers for complex traits. The relative stability of methylation profiles may enable better assessment of chronic exposures compared to single time-point protein measures. We present the first large-scale epigenetic study of the highly-abundant serum proteome measured via ultra-high throughput mass spectrometry in 14,671 samples from the Generation Scotland cohort. We further demonstrate the first large-scale comparison of protein EpiScores and their respective proteins as predictors of incident cardiovascular disease.

**Results** Marginal epigenome-wide association models, adjusting for age, sex, measurement batch, estimated white cell proportions, BMI, smoking and methylation principal components, reveal 15,855 significant CpG – protein associations across 125 of 133 proteins $P_{Bonferroni} < 2.71 \times 10^{-10}$. Bayesian epigenome-wide association studies of the same 133 proteins reveal 697 CpG-Protein associations (posterior inclusion probability > 0.95). 112 protein EpiScores correlate significantly with their respective protein in a holdout test-set. Of these, sixteen associate significantly with incident all-cause cardiovascular disease ($N_{events}$=191) compared to one measured protein.

**Conclusions** We highlight a complex interplay between the blood-based methylome and proteome. Importantly, we show that protein EpiScores correlate with measured proteins and demonstrate that the, as-yet understudied, high-abundance proteome may yield clinically relevant biomarkers. The protein EpiScores demonstrate more significant associations with cardiovascular disease than directly measured proteins, suggesting their potential as clinical biomarkers for monitoring or predicting disease risk. We suggest that biomarker development could be enhanced by the consideration of protein EpiScores alongside measured proteins.

**Keywords** Epigenetics, Proteomics, Cardiovascular disease, Biomarkers

## Background

Blood-based protein measurements are under increasing focus for the development of biomarkers of morbidity and mortality [1, 2], with evidence that protein prediction models outperform models using clinical information [2]. In addition to genetic influences, a complex interplay exists between the proteome and other omics layers, such as the methylome [3]. The methylome describes the pattern of genome-wide DNA methylation (DNAm), the addition of a methyl group to cytosine nucleotides, most commonly occurring at cytosines which precede a guanine (CpG sites) acting to regulate transcription [4]. Epigenome-wide association studies can help elucidate the relationship between DNAm and the circulating proteome, providing greater insight into the latter's regulation and the potential role of various environmental, biological and lifestyle factors on health.

The use of DNAm-based proxies for complex traits, including protein levels and health outcomes, represents an expanding field of research [5]. Where proteins are concerned, these proxies, which we call epigenetic scores (EpiScores), have been shown to display a more stable longitudinal measurement [6], likely reflecting cumulative and sustained impacts of environmental effects and biological changes [7]. This property is key to the promise of EpiScores as biomarkers and tools for risk prediction, stratification and precision medicine. Research thus far has demonstrated that EpiScores can track proteomic markers to enhance our understanding of the impact of chronic inflammation on both cardiovascular and neurological health [8, 9]; highlight novel associations with incident disease [10]; and augment prediction models for disease, offering measurable improvement in prediction over and above traditional risk factors [11]. Therefore, in the search for disease biomarkers to enhance risk prediction and monitor interventions to improve outcomes, both epigenetic and proteomic biomarkers should be considered.

Studies of the circulating proteome have predominantly captured lower abundant signalling and tissue-leakage proteins, using multiplexed antibody or aptamer-based assays, such as the Olink® and SomaScan platforms, which capture up to ~11,000 proteins. These targeted approaches navigate the challenges of the wide dynamic range of the human proteome and can quantify low-abundant proteins, many of which are potential biomarkers. However, quantification of high-abundance proteins, which constitute 99% of total protein mass in blood [12], using these assays is challenging due to the presence of multiple isoforms and saturation of affinity reagents, limiting the dynamic range of measurement [13]. In contrast to tissue leakage or signalling proteins, these proteins mostly function in processes occurring within blood, such as nutrient transport, innate immunity, or coagulation. Such proteins are well quantified using mass spectrometry (MS) [14] and data acquisition can be untargeted, independent of current paradigms or knowledge [12].

In this study, we conduct the first large-scale epigenome-wide association studies of the serum proteome as measured by mass spectrometry, to uncover novel associations between DNAm and proteins. Further, in independent data subsets, we train protein

EpiScores and then test their associations with incident cardiovascular disease. We then provide the first large-scale analysis of both measured proteins and protein EpiScores in association with incident cardiovascular disease. Figure 1 illustrates the project overview.

## Results

### EWAS

Marginal linear regression models were run using OmicS-data-based Complex trait Analysis (OSCA, v0.46.1) [15]. In these models each CpG was independently regressed on protein level, yielding 15,855 significant CpG – protein associations across 125 of the 133 uniquely mapped proteins ($P_{Bonferroni} < 2.71 \times 10^{-10}$), with a median genomic inflation factor of 1.17 (Additional file 1: Tables S1 & S2). These models do not account for inter-probe correlations or attempt to fine-map the findings. Consequently, we applied a joint and conditional Bayesian regression approach (Bayesian Grouped Mixture of Regressions Model for analysing OMIcs data, GMRMomi) [16, 17]. This resulted in 697 CpG – protein associations (PIP >0.95) for 120 of 133 uniquely mapped proteins, involving 457 unique CpGs. 286 of these were *cis* associations (CpG within 1 Mb of the

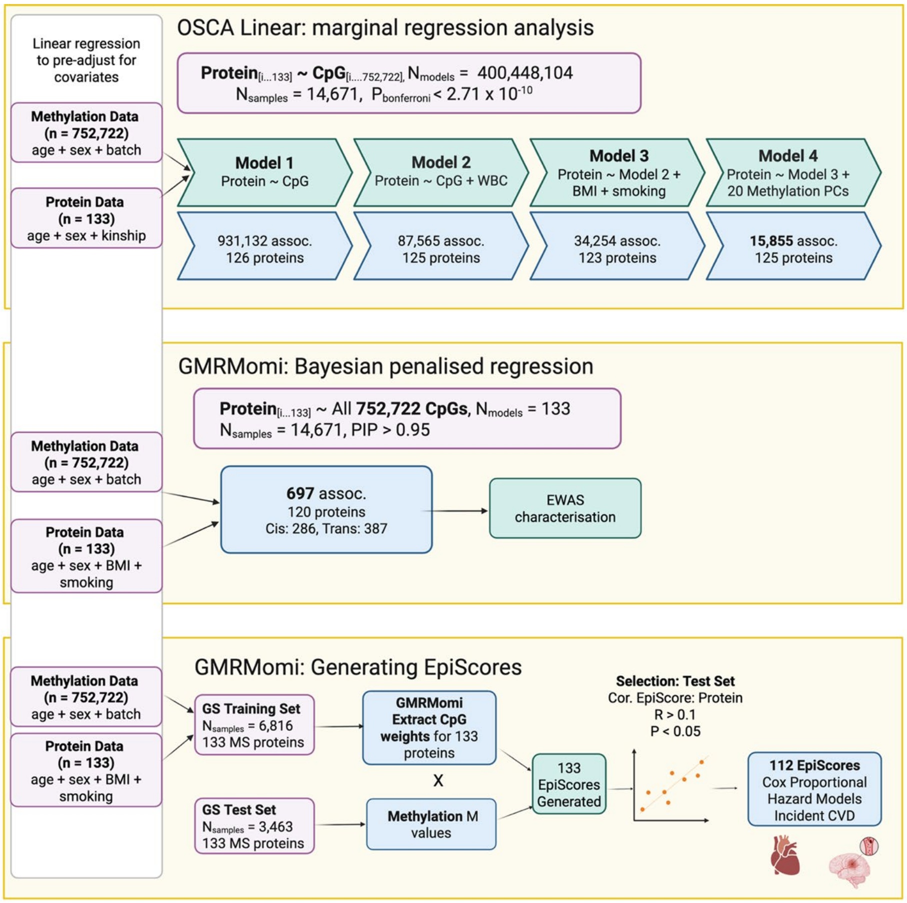
OSCA Linear: marginal regression analysis
Linear regression to pre-adjust for covariates
Methylation Data (n = 752,722) age + sex + batch
Protein Data (n = 133) age + sex + kinship
Protein[i...133] ~ CpG[i....752,722], Nmodels = 400,448,104
Nsamples = 14,671, Pbonferroni < 2.71 x 10-10
Model 1 Protein ~ CpG
Model 2 Protein ~ CpG + WBC
Model 3 Protein ~ Model 2 + BMI + smoking
Model 4 Protein ~ Model 3 + 20 Methylation PCs
931,132 assoc. 126 proteins
87,565 assoc. 125 proteins
34,254 assoc. 123 proteins
15,855 assoc. 125 proteins
GMRMomi: Bayesian penalised regression
Protein[i...133] ~ All 752,722 CpGs, Nmodels = 133
Nsamples = 14,671, PIP > 0.95
Methylation Data (n = 752,722) age + sex + batch
Protein Data (n = 133) age + sex + BMI + smoking
697 assoc. 120 proteins Cis: 286, Trans: 387
EWAS characterisation
GMRMomi: Generating EpiScores
Methylation Data (n = 752,722) age + sex + batch
Protein Data (n = 133) age + sex + BMI + smoking
GS Training Set Nsamples = 6,816 133 MS proteins
GS Test Set Nsamples = 3,463 133 MS proteins
GMRMomi Extract CpG weights for 133 proteins
X
Methylation M values
133 EpiScores Generated
Selection: Test Set
Cor. EpiScore: Protein
R > 0.1
P < 0.05
112 EpiScores Cox Proportional Hazard Models Incident CVD

**Fig. 1** Overview of EWAS and EpiScore workflow and results. For OSCA linear marginal regression analysis, each CpG is modelled individually for every protein within each model. For GMRMomi Bayesian penalised regression, all CpGs are modelled jointly. The Bayesian approach was subsequently used to identify lead CpGs and for the generation of protein EpiScores. WBC = estimated white blood cell proportions; BMI = log transformation of body mass index (kg/m$^2$); smoking = log transformation of smoking pack-years (+ 1); PCs = Principal Components; PIP = posterior inclusion probability. Created in BioRender, Marioni, R. (2025) https://BioRender.com/q80a293

transcription start-site (TSS) of the associated protein's gene) and 387 were *trans* associations (CpG more than 1 Mb outside of the associated protein's TSS or on a different chromosome) (Additional file 1: TableS3). 298 of the 457 CpGs map to 233 unique genes.

The Bayesian approach identified 505 of the protein-CpG associations (for 115 unique proteins and 333 unique CpGs) found with the OSCA approach (See Additional file 1: Table S4). The direction of the effect was consistent across all loci and there was a Pearson correlation of 0.94 in their effect sizes. Of the remaining 192 associations found via GMRMomi, 67 were present at $P < 3.6 \times 10^{-8}$ (epigenome-wide significance threshold) [18], and an additional 119 present at $P < 0.05$ in the OSCA analyses.

### Epigenetic architecture of the lead findings from the Bayesian EWAS

The distribution of associations by protein and CpG for the 697 lead findings from the Bayesian EWAS are displayed in Fig. 2. These demonstrate that most proteins (59%) have 6 or fewer associations, with a maximum number of 20 for P02750 (Leucine-rich alpha-2-glycoprotein). There was a strong concordance ($r = 0.64$) between the number of CpGs with PIPs >0.95 and the mean proportion of variance explained by all CpG loci for each protein, supporting the investigation of DNAm proxies for proteins (Fig. 2C). Most CpGs (79%) had one protein association, with up to 22 associations for one CpG (cg06072257, in open sea on chromosome 1, see Additional file 1: Table S5). Correlation analysis of the 22 associated proteins revealed Pearson r values between − 0.37 and 0.62 between pairs of proteins, with several demonstrating no intercorrelation (Additional file 1: Table S6). Network analysis with StringDB [19] revealed both functionally related and unrelated proteins, including those involved in the complement and coagulation cascades. This locus was associated with 11 different immunoglobulin components, 3 complement proteins, in addition to Ficolin-3, Angiotensinogen, Serotransferrin, Alpha-1-acidglycoprotein 1, Cell division cycle 5-like protein, C4b-binding protein alpha and beta chains and Vitronectin (Additional file 2: Figs. S1. and S2.). Across the 133 Bayesian protein EWASs, the significant loci were not distributed proportionally across the genomic regions (e.g. OpenSea, CpG Islands etc.) captured by the array ($X^2$: 62.99, df = 5, $P = 2.92 \times 10^{-12}$, Fig. 2D). For example, there was an enrichment of findings within OpenSea regions but fewer findings than would be expected in CpG Islands. The majority (56%) of the effect sizes were identified in *trans* locations although there were no differences by direction or magnitude between *cis* and *trans* associations (Fig. 2E). The distribution of *cis* and *trans* associations across the genome is displayed in Fig. 2F.

### EWAS catalogue

The EWAS catalogue [20] was searched (download date: 05/08/25) to identify any previously reported associations for our 697 results from the Bayesian model. After filtering to entries from "whole blood" and $P < 3.6 \times 10^{-8}$ and matching on UniProt ID, the protein gene or the word "protein", four previously identified associations were identified - all from our previous EWAS of SomaScan proteins [21] (Additional file 1: Table S7). That study ($n_{individuals} = 774$, $n_{proteins} = 4,058$) [21] featured 60 unique proteins that were also present in the MS dataset. There were 12 CpG loci with $P < 3.6 \times 10^{-8}$ for 8 of these 60 proteins, of which four (associating with three proteins) had a PIP >0.95, all with concordant effect size directions in our current analyses. If the P-value threshold is relaxed to < 0.05, there are 4107 CpG loci demonstrating associations with the 60 proteins, of which

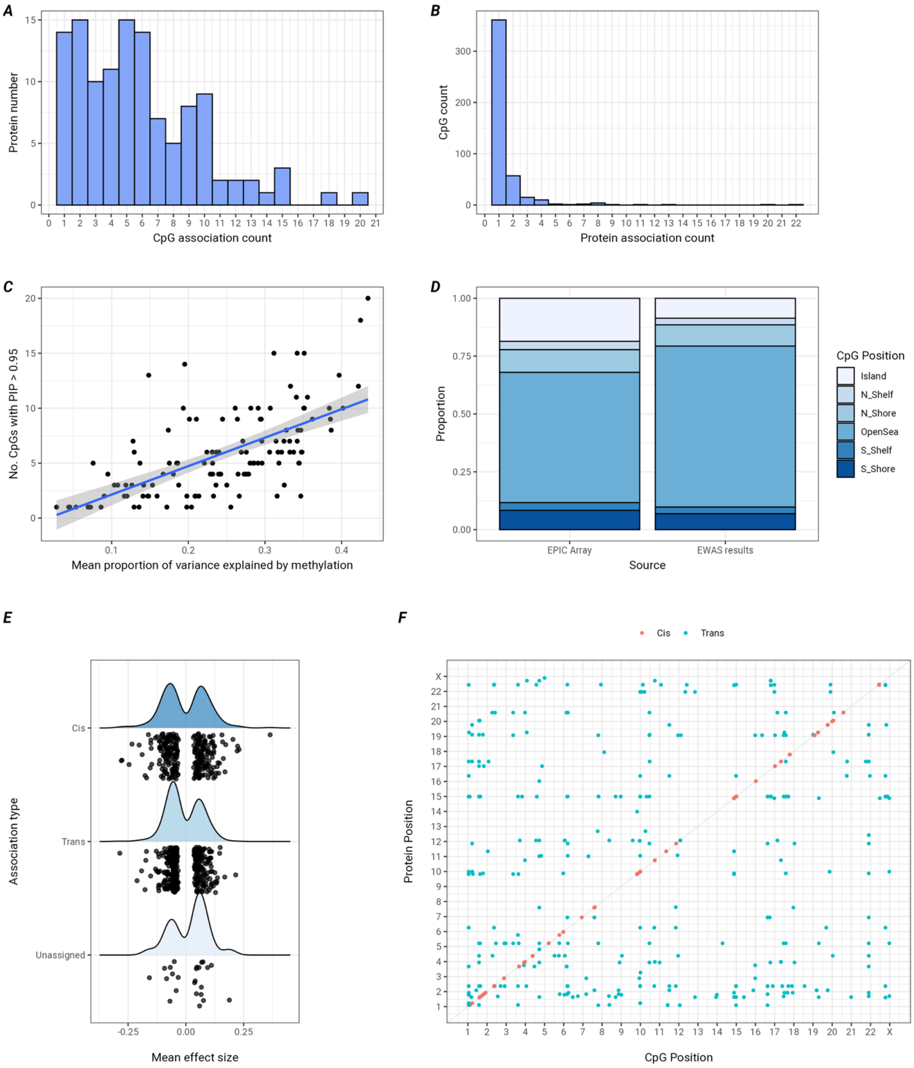

**Fig. 2** Summary of 697 protein ~ CpG associations from the Bayesian EWAS results. **A** The distribution of number of proteins by number of CpG associations; **B** The distribution of number CpGs by number of protein associations; **C** The correlation between the number of CpG association of each protein, by the mean proportion of variance explained by all CpG loci; **D** The proportion of CpGs in regions, specified by relation to CpG islands for the EPIC array and for the Bayesian EWAS results, demonstrating enriched results in Open Sea and reduced in Island regions; **E** Mean effect size of associations by association type, demonstrating the effect size is similar whether the association is in cis or trans. Unassigned associations are those for which the protein gene could not be annotated to a position in GRCh37 (*N* = 24, Additional file 3: Methods M3); **F** Each association plotted by genomic position of the protein gene and CpG probe demonstrating the distribution of associations across the genome

24 had a PIP >0.95 in the current study, with a correlation of effect sizes of 0.8 ($P$ = 2.6 x $10^{-6}$). It is important to note the differences between the methods employed by the two studies (marginal regression vs. joint and conditional modelling of the CpGs). Further, the SomaScan EWAS considered plasma proteins rather than serum as in the current MS proteome analysis. Although both studies utilised data from Generation Scotland, the biosamples were taken at different time points, with the blood samples for plasma being obtained between 2015 and 2018, compared to between 2006 and 2011 for serum samples analysed via MS [22].

To identify any traits previously associated with the 457 lead loci, we again filtered to entries from “whole blood” with $P < 3.6 \times 10^{-8}$ and further to studies with $N$ >1000.

231 of the 457 CpGs have previous trait-associations documented in the EWAS catalogue, 105 of these with more than one trait (see Additional file 1: Table S8). This demonstrates that our results align with and expand on previous associations identified in other EWASs. For example, we identified additional and relevant protein associations of cg00574958 (*CPT1A* gene) and cg06500161 (*ABCG1* gene), previously associated with multiple metabolic traits [23, 24]. Similarly, we identified relevant protein associations for cg19693031 (*TXNIP* gene) previously associated with type 2 diabetes [23, 25] and cg07839457 (*NLRC5* gene) previously associated with immune-related proteins such as CD48 antigen [26] or traits such as rheumatoid arthritis [23].

## EpiScores

We generated 133 protein EpiScores using the Bayesian GMRMomi approach which, through joint and conditional modelling of all CpG loci provides a parsimonious solution for each protein. The number of CpGs with non-zero weights for each EpiScore, the majority of which (91%) can also be found on the EPICv2 array, are summarised in Additional file 1: Table S9.

Of the 133 Protein EpiScores, 112 had Pearson $r > 0.1$ and $P < 0.05$ with rank-based inverse normalised proteins when projected into an independent Generation Scotland test set ($n = 3{,}463$) (Fig. 3., Additional file 1: Table S10). These patterns largely persisted when we reprojected the EpiScores using loci common to both EPICv1 and EPICv2 arrays (Additional file 1: Table S11, Additional file 2: Fig. S3).

The 112 EpiScores and their corresponding proteins were then studied in relation to incident cardiovascular disease via Cox proportional hazards models with a follow-up duration of up to 17.6 years ($n = 3{,}345$, after excluding those with a prevalent cardiovascular disease diagnosis – see Methods) (Fig. 4).

We compared the number and magnitude of statistically significant associations (hazard ratios) for both the protein EpiScores and directly measured proteins in relation to incident cardiovascular disease. After adjustment for age and sex, 16 protein EpiScores demonstrated a significant relationship with the composite cardiovascular disease outcome, compared to only one of the measured proteins ($P_{\mathrm{Bonferroni}}$ <

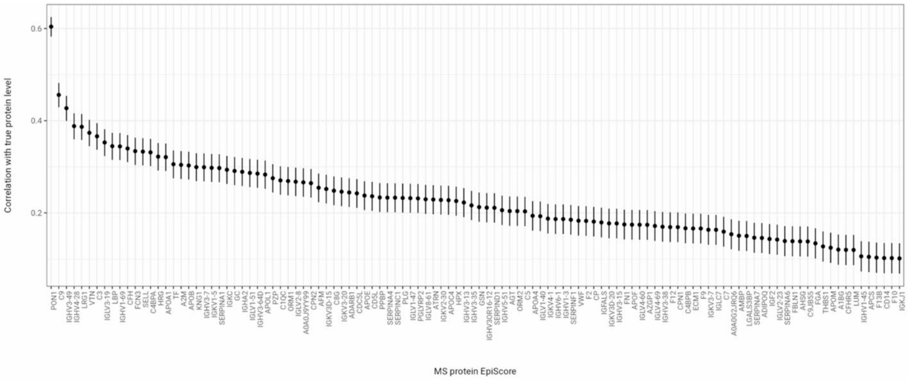

**Fig. 3** Pearson correlation of 112 EpiScores and proteins in the Generation Scotland test set. Test set N = 3,463. Correlation results displayed for 112 EpiScores where Pearson *r*>0.1 and *P* < 0.05 using the EPICv1 loci. Central dot represents Pearson r and the error bars represent 95% confidence intervals. Proteins are labelled by gene, except for Ig-like domain-containing protein 1 (A0A0G2JRQ6) and 2 (A0A0J9YY99), annotated by UniProtID. These proteins were annotated to scaffolds or patches in build hg19 and have not been assigned gene names (see Additional file 3: Methods M3.). Transferrin (C9JB55, 75 amino acids) is also labelled by UniProtID as it originates from the same gene as Serotransferrin (P02787, 698 amino acids, labelled TF)

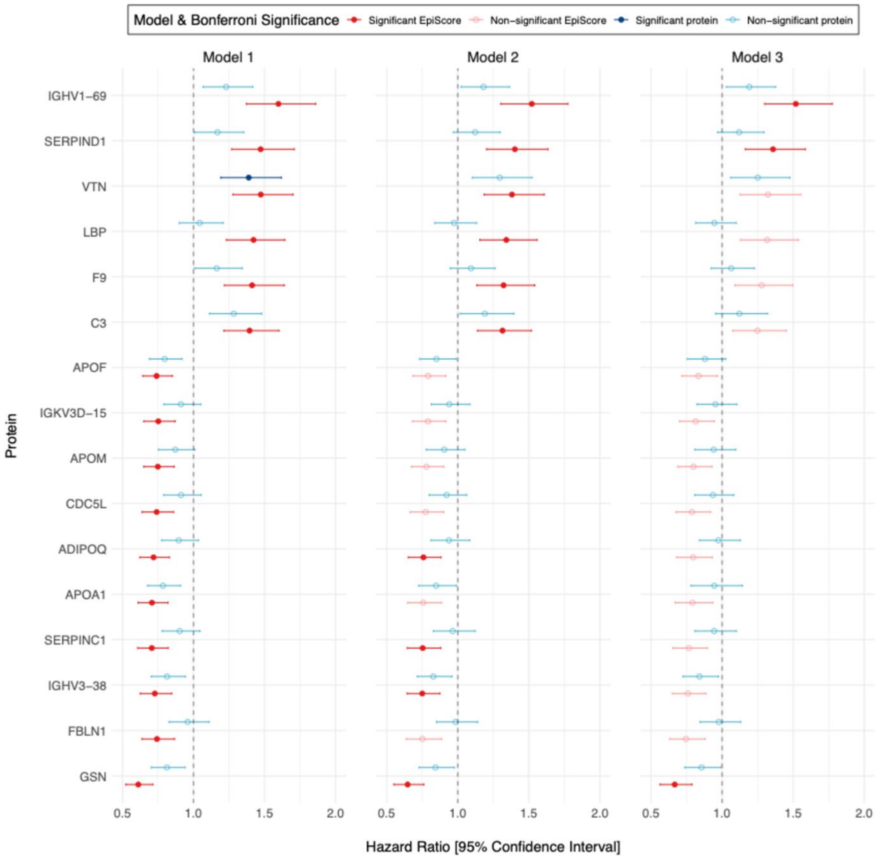

**Fig. 4** EpiScore and measured protein hazard ratios for time-to incident cardiovascular disease. Results are displayed where either protein or EpiScore demonstrate Bonferroni-significant associations ($P < 0.05/112$) in model 1. Model 1: TTE ~ EpiScore/Protein + age + sex; Model 2: TTE ~ EpiScore/Protein + age + sex + BMI + smoking + alcohol; Model 3: TTE ~ EpiScore/Protein + age + sex + BMI + smoking + alcohol + diabetes + hypertension + HDL cholesterol + Total cholesterol + average systolic blood pressure + average diastolic blood pressure. EpiScore/Protein denotes EpiScore or protein as a predictor variable. HR = Hazard Ratio per SD of the predictor, CI = 95% confidence interval. Colour in bold denotes significance at $P_{\text{Bonferroni}} < 4.46 \times 10^{-4}$ (= 0.05/112). Proteins are labelled by gene, with the exception of Ig-like domain-containing protein 1 (A0A0G2JRQ6) and 2 (A0A0J9YY99), annotated by UniProtID, which were annotated to scaffolds or patches in build hg19 and have not been assigned gene names (see Additional file 3: Methods M3.). Transferrin (C9JB55, 75 amino acids) is also labelled by UniProtID as it originates from the same gene as Serotransferrin (P02787, 698 amino acids, labelled TF)

$0.05/112 = 4.46 \times 10^{-4}$); 52 EpiScores and 21 proteins – mapping to 56 unique proteins – were significant at a nominal $P < 0.05$ threshold (Additional file 1: Tables S12 and S13). Furthermore, the absolute values of the log-hazards (per standard deviation of the predictor variable) were consistently greater for the EpiScores compared to the corresponding measured proteins (Fig. 4., full results: Additional file 1: Table S13 and Additional file 2: Fig. S4). These findings remain nominally significant ($P < 0.05$) upon further adjustment for covariates relevant to cardiovascular disease. The proportional hazards assumption was met for all models (Schoenfeld residual $P_{\text{global}}$>0.05, $P_{\text{EpiScore/Protein}}$ >0.05, Additional file 1: Tables S14 & S15) with the exception of one EpiScore (Cell division cycle 5-like protein, CDC5L, Q99459) and one protein (Vitronectin, VTN, P04004).

Finally, we compared nested Cox proportional hazard models to determine if the EpiScore augmented the measured protein in the incident cardiovascular disease analyses. Here, we considered the 17 instances where both the model with the EpiScore and the model with the corresponding, directly measured protein had nominally significant

($P < 0.05$) associations for the EpiScore and protein. Adding the EpiScore to a model controlling for age, sex and the measured protein resulted in a significantly improved model fit for 10 of the 17 models at $P < 0.003$ (0.05/17), (Additional file 2: Fig. S5., and Additional file 1: Tables S16 and S17).

## Discussion

Here, we present the first large scale epigenome-wide assessment of the highly abundant serum proteome, as measured by mass spectrometry. Using two different regression frameworks, we revealed 15,855 significant associations via a marginal linear approach and 697 associations using a penalised Bayesian approach. 505 of the protein ~ CpG associations overlapped between the two methods. The two EWAS approaches offer different insights into the relationship between the methylome and the proteome. As the marginal regression models consider each CpG in isolation, ignoring any correlation patterns across the genome, it identifies a large number of associations. By contrast, the penalised Bayesian approach implicitly performs fine mapping through the joint and conditional analysis of all CpGs. This yields parsimonious solutions for downstream analyses and EpiScore applications [16].

As we have shown for other complex traits, such as markers of metabolic health [27], the method used to conduct an EWAS has major implications on the number of significant loci identified. Here, a marginal approach identified 15,855 lead loci, compared to 697 in the joint and conditional penalised Bayesian regression. Both methods offer valid insights. The former highlights associations for intercorrelated CpGs, potentially across genes which are functionally related. The latter identifies a parsimonious set of high-confidence lead loci, reducing the chance of false-positive findings.

Focussing on the Bayesian EWAS, our results include both previously unreported associations and those that align with and expand upon results reported in the literature. For example, some of our loci, mapping to *ABCG1*, *CPT1A* and *TXNIP*, have been associated with lipid and metabolic traits such as triglycerides, BMI, and type 2 diabetes [23–25, 28, 29]. Four CpGs within *ABCG1*, which encodes a protein involved in cholesterol transport [30], associated with 12 different proteins. For example, cg06500161 had 8 *trans* associations with apolipoprotein F, apolipoprotein C4, afamin, gelsolin, vitronectin, apolipoprotein A-I, antithrombin-III, and plasminogen. Additionally, cg00574958 (*CPT1A*, important for mitochondrial oxidation of long-chain fatty acids [31]), was found here to be associated with Apolipoprotein E, Apolipoprotein B-100, and two immunoglobulin components, complementing previous protein associations with APOC3 and CRP [21, 32]. We identified associations for cg19693031 (*TXNIP*, an oxidative stress mediator, particularly associated with diabetes [33]) and eight different proteins, including those involved in the innate immune response such as Ficolin 3, Attractin and Complement component C9, which provides further insight into the link between diabetes and inflammation [34]. Finally, we build on previously identified associations between CpGs annotated to *NLRC5* and immune-system proteins [26]. We found associations between cg0783945 (*NLRC5*) and seven proteins, including four immunoglobulin components, 2 complement proteins and plasminogen whilst cg16411857 (*NLRC5*) was also associated with an immunoglobulin component, thus reinforcing the role of the *NLRC5* as a regulator of immune responses [35].

In our results, the most pleiotropic CpG site (cg06072257) associated with 22 proteins. The CpG itself is not directly annotated to a gene, but is situated near *UBIAD1*, which encodes UbiA Prenyltransferase Domain Containing 1, a protein involved with antioxidant processes [36]. It has previously been associated with prevalent breast cancer [23], incident COPD [23], an age-by-sex interaction [37] and the levels of 17 proteins including CRP, SERPING1, CHAD and CGA [21, 32]. However, none of these proteins previously identified were assessed in our dataset. A network analysis using StringDB [19] highlighted multiple interactions (Additional file 2: Fig. S2). For example, complement 8 alpha and gamma chains and vitronectin are classified as having 'known interactions' (KEGG annotated pathway: Complement and coagulation cascades, via StringDB). However, in our data, vitronectin did not correlate strongly with the complement 8 alpha or gamma chains ($r_{Pearson}$ = −0.07, 0.017, respectively) and only shared this one high confidence CpG locus. Further, Ficolin-3 and complement factor H, which were weakly correlated ($r_{Pearson}$ = 0.16, S17), are both annotated to the serine-type endopeptidase complex (Protein complexes annotated by the Gene Ontology Consortium, as of August 2022, via StringDB). The association of this CpG locus with proteins of related functions, in our results and previous EWASs [21] suggest it could play a regulatory role or reflect changes in the activity of these pathways. However, this analysis was partly limited by incomplete recognition of genes for immunoglobulin proteins in the StringDB database and further experimental evidence is required to explore this in detail.

In independent train/test Generation Scotland subsets, we generated statistically significant EpiScores for 112 proteins using CpGs from the EPICv1 array. As suggested from our sensitivity analyses, not all will translate perfectly to different Illumina array versions. Sixteen of the generated EpiScores were associated significantly with incident cardiovascular disease compared to just one measured protein, in an age and sex adjusted model. In general, the hazard ratio estimates for EpiScores were either higher (where $HR > 1$) or lower (where $HR < 1$) than the respective measured protein estimates. Further, it is not necessarily the EpiScores which correlate the highest with their respective protein which demonstrate the strongest disease associations. For example, Gelsolin (GSN, P06396) associates with cardiovascular disease with a hazard ratio of 0.6, $P = 4.9 \times 10^{-10}$ and demonstrates a Pearson r of 0.21 with its paired measured protein. This pattern extends to models adjusted for cardiovascular-relevant covariates, where three EpiScore associations remained Bonferroni significant.

Many of the EpiScore findings align with previously studies on proteins and cardiovascular disease. For example, vitronectin (P04004) has already attracted considerable interest as a putative biomarker for cardiovascular disease. It is a glycoprotein found both in blood, alpha granules of platelets and the extracellular matrix, with suspected roles in platelet aggregation following vascular injury and, via plasminogen activator inhibitor-1, a role in reducing thrombus clearance [38]. Vitronectin levels have been shown to correlate with the extent of coronary atherosclerosis [39], to be higher in patients with acute coronary syndromes [40] and be an independent risk factor for adverse cardiovascular events in patients undergoing percutaneous cardiac interventions [41]. Thus, it has been suggested as a biomarker to increase the accuracy of acute coronary syndrome diagnosis [42]. Our results ($P04004_{EpiScore}$ HR per SD (95% CI): 1.48 (1.28,1.7), compared to $P04004_{measured}$ 1.39 (1.19, 1.62)), which focused on longer term cardiovascular disease prediction, suggest it would be worth considering an EpiScore for vitronectin alongside the measured protein itself when investigating biomarker utility. Similarly, lipopolysaccharide-binding protein (P18428) has also been associated with increased risk of cardiovascular disease [43], and the EpiScore demonstrated a stronger association with incident cardiovascular disease 1.42 (1.23–1.64) compared to 1.04 (0.9–1.2) for the measured protein.

By contrast, Heparin cofactor 2 (P05546, SERPIND1) is a thrombin inhibitor, and has predominantly been suggested to be protective against cardiovascular disease [44, 45]. However, our findings showed an increased risk of a cardiovascular diagnosis or death with increasing concentrations (EpiScore HR 1.47 [1.27–1.71], measured protein 1.17 [1.01–1.35]).

Gelsolin (P06396, GSN) is another interesting example. This protein facilitates actin filament recombination and circulating GSN is proposed to mitigate the development of atherosclerosis through multiple pathways including inflammatory cell migration and interleukin release and limiting endothelial injury [46]. In our results, the GSN EpiScore demonstrated a significant protective association with incident cardiovascular disease (HR: 0.61 [0.52,0.71], $P = 4.96 \times 10^{-10}$), whilst the measured protein demonstrated a protective, but non-significant relationship after correction for multiple comparisons (HR: 0.81, $p = 0.005$).

Blood-based DNAm measures methylation predominantly in leucocytes [47]. High abundant proteins, comprising up to 99% of plasma proteins by mass [12] are mostly synthesised in the liver and secreted into plasma [48]. Therefore, the association between blood-based DNAm measurements and protein abundance is indirect. Nevertheless, we have demonstrated that blood-based DNAm significantly associates with relative protein abundance in 112 of 133 proteins assessed. We hypothesise that DNAm patterns reflect biological changes underpinning variation in these proteins, largely influenced by factors impacting protein regulation and synthesis such as inflammatory and metabolic states.

For example, ApoA-1 is a major component of high-density lipoproteins, transporting cholesterol from tissues to the liver for excretion, with additional immuno-modulatory and anti-inflammatory roles [49]. Synthesis of ApoA-1 occurs principally in the liver and also the intestine, regulated by hormonal mediators such as oestrogen, thyroid hormone and insulin [50], in turn impacted via lifestyle factors such as diet and exercise [51]. Further, systemic inflammation, a feature of metabolic syndrome, can inhibit Apo-AI production via inflammatory cytokines such as IL-6 and TNF-alpha [52]. There is a well-established impact of inflammation and metabolic state on DNA methylation in blood [9, 32] and thus it is likely these common features underpin correlations between EpiScores and proteins. Multiple studies have identified an association of ApoA-1 or ApoA-1:lipid ratio and cardiovascular disease [53, 54]. We found Apolipoprotein A1 (ApoA-1) was associated with decreased hazards of cardiovascular disease ($HR_{\text{measured protein}}$:0.79, $P = 0.001$, $HR_{\text{EpiScore}}$: 0.71, $P = 4.76 \times 10^{-06}$) in the age and sex adjusted model.

Though the strength of the mass spectrometry approach is its untargeted nature, reducing potential selection bias, it does not easily measure low abundance proteins. As approximately 20 proteins constitute 98% of the protein mass in human plasma, mass spectrometry may lose valuable information from low-abundance proteins which remain highly relevant for health and disease, unless particular strategies are employed to allow their accurate detection [55]. An example of this could be interleukins, which occur at concentrations of ng/L in contrast to, for instance, apolipoproteins at concentrations of g/L [56].

Alternative mass spectrometry approaches such as the Seer platform [57] have also been used to develop protein EpiScores (termed epigenetic biomarker proxies). As further test-sets become available, it will be interesting to compare the performance of EpiScores for the same protein generated on different platforms.

As the Generation Scotland cohort is predominantly of white European ancestry and limited to those living in Scotland, these results are not necessarily generalisable to other populations, although previous work has demonstrated that EpiScores for diabetes risk, metabolic traits, CRP and smoking have all translated well to other cohorts, including those of diverse ancestries [11, 27, 32, 58]. As the methylation and protein data are cross-sectional, results and interpretation could be strengthened by analysis in longitudinal datasets, allowing within subject comparison of changes in methylation and protein levels. Additionally, DNAm is only one form of epigenetic regulation. To form a complete understanding of the relationship between the epigenome and the circulating proteome we could also consider additional factors such as histone modification, chemical modification of RNA and mitochondrial gene expression [7].

Our analyses demonstrates that protein EpiScores exhibit significant relationships with incident cardiovascular disease where measured proteins do not and tend to demonstrate stronger hazard ratio point estimates. However, further modelling, in a larger dataset, including other known cardiovascular disease risk factors should be undertaken to probe this relationship further.

## Conclusions

Here, we conducted the first large scale methylome-wide analyses of the high abundant serum proteome as measured by mass spectrometry. Using two separate statistical frameworks, we identified 505 common CpG-protein associations, the majority of which were *trans* associations. We find a complex interplay between these omics layers, including a single CpG (cg06072257) associating as a *trans* locus with 22 proteins, whilst most CpGs are associated with a small number of proteins. Furthermore, we generated EpiScores for 112 proteins, 16 of which demonstrated both significant and stronger associations with incident cardiovascular disease compared to the respective measured proteins. There was also evidence for additive effects when including both the measured protein and its corresponding EpiScore in the same model.

Our results demonstrate the potential for protein EpiScores as disease biomarkers, particularly applicable to non-communicable diseases associated with environmental and lifestyle factors. The potential for applications in disease prediction, evaluating therapeutic intervention, risk stratification and precision medicine warrant further investigation alongside proteome measurements.

## Methods

### Generation Scotland

Generation Scotland is an epidemiological study with comprehensive DNA, clinical, and socio-demographic data from approximately 24,000 volunteers with linkage to medical records [59]. Participants were recruited from across Scotland, aged 17–99, between 2006 and 2011. Blood samples were taken during the initial clinic visit for just over

20,000 volunteers, alongside health, cognitive and lifestyle questionnaires. Participants provided informed consent to electronic-health records linkage to both secondary and primary care data, allowing analysis of prevalent and incident disease.

### Mass spectrometry proteomics

Measurement of the circulating proteome in Generation Scotland was carried out using a high flow-rate liquid chromatography tandem mass spectrometry, using SWATH acquisition [60]. Data processing was performed with DIA-NN, using a spectral library approach [61].This generated data for 439 inferred proteins, for 15,818 participants at the time of analysis.

Full details of the protocol have been described previously [62]. In brief, serum samples were pre-processed for protein denaturation and trypsinisation, prior to liquid chromatography - mass spectrometry (LC)-MS, using the Agilent 1290 Infinity II system and TripleTOF 6600 mass spectrometer (SCIEX) and a scanning SWATH method [60]. Output data were processed by DIA-NN [61], identified using a spectral library [63] with precursor false discovery rate (FDR) set to 1%. R was used for further post-processing including within-batch drift correction using a previously described method [64] and between-batch correction, using the "limma" v3.54.2 algorithm [65]. Identified signals were mapped to Universal Protein Resource (UniProt) IDs [66]. 133 of the signals were uniquely mapped to one protein (Additional file 1: Table S18) and the remaining 306 were mapped to multiple possible target outcomes (Additional file 3: Methods M1). For computational efficiency, we focus here on the 133 individual proteins, the values for which were rank-based inverse normalised before being taken forward for further analysis.

### DNA methylation

Whole blood-based DNAm measurements were profiled on samples from the Generation Scotland baseline appointment, on sodium bisulphite treated DNA, with the Illumina Infinium HumanMethylationEPIC BeadChip array v1.0 [67]. Sample processing and quality control (QC) of the methylation data is described in Additional file 3: Methods M2, and has previously been described in full [22]. Briefly, DNAm was profiled in four separate sets (Post QC: $N_{Set1}$ = 5,087, $N_{Set2}$ = 459, $N_{Set3}$ = 4,450, $N_{Set4}$ = 8,873, Total $N$ = 18,869) [22]. Samples were removed if the median methylated signal intensity was over three standard deviations lower than the expected value, where the methylation-derived sex differed from self-reported sex and if >0.5% CpGs in the sample had a detection $P$-value >0.01 (or >1% of CpGs with a detection $P$-value >0.05 for set 1). Poorly performing probes were also removed if the beadcount was < 3 in >5% of samples or >1% of the samples had a detection $P$-value of >0.01 (>0.5% samples with detection $P$-value >0.05 in set 1). A total of 752,722 CpG sites were included following quality control. Normalised methylation M-values were used for downstream analyses. 14,671 individuals from the Generation Scotland cohort had complete methylation and protein data for analysis.

### Epigenome-wide association studies (EWASs)

EWASs were performed to identify CpG-protein associations. We first employed a marginal linear model approach (separate linear model for each possible CpG-protein association), using OmicS-data-based Complex trait Analysis (OSCA, v0.46.1) [15]. Marginal regression analyses do not account for the correlation structure between CpGs across the genome, considering each locus in isolation. This approach is most commonly taken in omics association studies but can lead to issues with genomic inflation [68] which was observed here. We subsequently employed a newly developed Bayesian approach (GMRMomi) which models all CpGs jointly and conditionally on each other for each protein [16]. This estimates CpG effects whilst considering relationships between probes, selecting highly influential probes amongst those which are correlated. Therefore, whilst the marginal linear analysis seeks to identify all CpGs associated with the phenotype of interest, the penalised Bayesian approach facilitates both dimensionality reduction and fine mapping to identify the most strongly influential probes and parsimonious EpiScore signatures. In order to reduce the computational burden of the >400 million linear and 133 Bayesian EWASs, methylation M-values were pre-regressed for age, sex and measurement batch using the limma package (version 3.60.4) in R [65]. Residuals from the output of a linear model for each CpG were scaled to have a mean of zero and unit variance prior to the EWASs. Demographics for the 14,671 included individuals can be found in Additional file 1: Table S19. All analyses outside of OSCA and GMRMomi were conducted in R version 4.4.1.

### OSCA linear

Rank-based inverse normalised values for 133 proteins were taken forward to mixed-effects linear regression analyses that adjusted for age and sex as fixed effects and a kinship matrix as a random effect using the lmekin function (coxme package, version 2.2.20 [69]). The kinship matrix accounts for relatedness within the known family structures in Generation Scotland. Residuals from each model (one per protein) were then scaled to have mean of zero and unit variance prior to downstream analysis.

The fast-linear option within OSCA was used for the frequentist EWASs, with an iterative approach for including covariates known to impact methylation [70], see Fig. 1. Added covariates included: estimated white blood cell proportions (Houseman method [71], with neutrophils dropped to minimise collinearity) in model 2. In model 3, we further included body mass index (BMI, kg/m$^2$) and smoking pack-years, where one smoking pack-year equates to smoking 20 cigarettes per day for one year, both known to affect DNA methylation [27, 58]. Both variables were log transformed (a constant of 1 was added to all values in the smoking variable to account for never smokers in the transformation), to minimise skew in the data. Missing data (BMI: $n$ = 92, smoking pack-years, $n$ = 275) were mean-imputed using the impute_mean function from the missMethods package, version 0.4.0 in R [72], prior to log transformations. Finally, in model 4, we added the first 20 principal components of the methylation data to account for possible unmeasured confounding, given ongoing evidence of model inflation (Additional file 2: Fig. S6) and as previously employed in EWASs in GS [22, 27]. The relationship between the principal components and other continuous covariates is illustrated in Additional file 2: Fig. S7. The strongest correlation observed were those between PC6 and eosinophil proportion and PC9 and eosinophil proportions (Pearson $r$ = 0.25 and $r$ = −0.25, respectively, Additional file 1: Table S20). The strongest correlations observed between proteins and PCs were 0.09 (PC14 and C3, P01024) and − 0.09 (PC15 and PON1, P27169), see Additional file 1: Table S21. A Bonferroni-corrected threshold of $P_{\text{Bonferroni}} < 2.71 \times$

$10^{-10}$ was set for statistical significance. This was calculated using $P < 3.6 \times 10^{-8}$ as the epigenome-wide significance threshold divided by the number of proteins assessed (133) [18].

### GMRMomi

GMRMomi is a software implementation of Bayesian penalised regression [16, 17] based upon a framework proposed for genomics data (GMRM), adapted for large-scale multi-omics data. The method utilises Gibbs sampling to generate draws from the posterior distribution, considering the underlying genetic architecture and intercorrelation of CpG sites, modelling all CpGs jointly and conditionally on each other. In addition to controlling for known covariates, this method implicitly controls for unknown variables such as white-cell proportions, which would usually be estimated from the methylation data itself. This contrasts with the marginal analysis undertaken by OSCA, which considers each probe separately.

Methylation M-values were prepared as above. Rank-based inverse normally transformed protein levels were regressed on age, sex, logarithmic transformation of BMI, and the logarithmic transformation of smoking pack-years (+ 1). Any missing data (BMI: $n = 92$, smoking pack-years: $n = 275$) were mean imputed. The residuals from these linear models were scaled (mean zero, unit variance) and taken forward for further analyses. Prior mixture variance proportions were set to 0.0, 0.001, 0.01, and 0.1, equivalent to negligible, small, medium and large CpG effect sizes, as previously used for Bayesian EWAS studies of the circulating proteome [73]. 2000 model iterations were run for each protein, with 750 'burn-in' iterations discarded prior to averaging the effect sizes over the last 1250 posterior samples. A posterior-inclusion probability (PIP) of >0.95 was used to select robust CpG-protein associations (Fig. 1). The mean proportion of variance explained by all CpG loci for each protein was calculated as the mean variance explained by methylation probes divided by total variance across the last 1250 iterations.

### Annotation of proteins and methylation sites

CpG sites were annotated using the minfi package in R (*IlluminaHumanMethylationEPICanno.ilm10b4.hg19)*, version 1.50.0 [74], to establish chromosome, probe position, relation to CpG-island and any nearby genes. Protein gene annotations were performed in R using BioMaRT and Ensembl (Genome reference consortium build37, GRCh37, to establish chromosome and transcription start-site (TSS) [75, 76].

CpG sites were characterised as being in *cis* (within 1 Mb) or *trans (*outside of this region or on a different chromosome) of the TSS of the associated protein's gene. It is important to note that these annotations are for positional and descriptive purposes, irrespective of the primary tissue responsible for protein synthesis and we make neither a directional nor causal assumption about the nature of the relationships between CpGs and proteins.

Three proteins could not be fully annotated to a genomic position in GRCh37, for further details see Additional file 3: Methods M3. A Chi-squared test and post hoc Z-tests assessed whether the distribution of CpG location (e.g., part of a CpG island) for the significant EWAS loci, differed from the distribution of all probes on the array.

Further analysis included an EWAS catalogue [20] search for previous associations for CpGs identified as statistically significant in the EWASs, along with the characterisation of pleiotropic loci and assessment for pathway enrichment in probe-associated genes (Additional file 3: Methods M4 & M5).

### Protein episcores

The 14,671 Generation Scotland participants were split into training and test datasets to build protein EpiScores. The training dataset contained 6,816 individuals from measurement Sets 1, 2 and 4. To minimise overfitting, all individuals ($n$ = 3,671) within the same family pedigree [77] as participants in the test dataset were removed from the training dataset. The test dataset contained 3,463 unrelated individuals who had DNAm processed together in Set 3. Demographics for the training and test sets can be found in Additional file 1: Tables S22 & S23. In the training dataset, GMRMomi was run as previously specified (prior mixture variance proportions were set to 0.0, 0.001, 0.01, and 0.1, total iterations to 2000 and 750 burn-in iterations. The mean posterior CpG weights, calculated from post-burn-in iterations, for each of the 133 protein regression models were extracted. EpiScores were then projected into the test set to create protein EpiScores (additive sum of all CpG weights multiplied by the measured CpG M-values). We additionally calculated the proportion of CpG loci for each EpiScore which can be found on the EPICv2 array [78] (Additional file 1: Table S9) and re-calculated the protein EpiScores using only the loci on both the EPICv1 and EPICv2 arrays. To gauge the translatability of these EpiScores to data where methylation has been measured using the EPICv2 array we assessed the correlation between both sets of EpiScores and their paired measured protein (Fig. 3., Additional file 1: Table S11).

### Protein episcore versus measured proteins: associations with incident cardiovascular outcomes

EpiScores which demonstrated a significant correlation with the measured protein level (Pearson $r$ >0.1 and uncorrected $P < 0.05$), were taken forward for further analysis in the test dataset ($N_{\text{individuals}}$ = 3,463). We explored the association of the protein EpiScores with incident cardiovascular disease using Cox proportional-hazards models [79]. Incident cardiovascular disease was defined as a composite outcome, including a diagnosis of coronary heart disease, ischaemic stroke, myocardial infarction and any death related to cardiovascular disease. Diagnoses were determined from secondary care records, using CALIBER/HDRUK consensus definitions [80] and cardiovascular disease related death using ICD codes I00-99, aligning with previous work on cardiovascular disease carried out in Generation Scotland [81] (for further details see Additional file 3: Methods M6). The censor date was set to the most recent date of linkage to disease data (August 2023) or non-CVD-related death, with a total follow up period of up to 17.6 years from baseline appointment ($N_{\text{CVD-diagnoses and deaths}}$ = 191, $N_{\text{censored}}$ = 3154). Any individuals with a diagnosis of cardiovascular disease received prior to their baseline appointment, were excluded from the analysis ($N_{\text{prevalent}}$ = 116). For baseline demographic and outcome details see Additional file 1: Table S24.

Within the Generation Scotland test dataset, Cox proportional hazards models were run for each protein EpiScore and each mass spectrometry measured protein level and time-to incident cardiovascular disease using the survival R package, version 3.7.0 [82]. To compare the strength of association of both EpiScores and their paired, measured proteins with incident cardiovascular disease we ran models adjusting for age and sex.

This approach has also been taken with other work assessing the potential of epigenetic biomarker proxies [57]. We additionally ran sensitivity analyses to control for further covariates: firstly BMI, smoking pack-years and alcohol consumption (units per week) (model 2) and finally also adding prevalent diabetes, prevalent hypertension, HDL cholesterol, Total cholesterol and average systolic and diastolic blood pressures (model 3). Missing data (BMI = 16, smoking pack-years = 4, alcohol units/week = 254, average systolic and diastolic blood pressures = 3, HDL cholesterol = 25 and total cholesterol = 21) were imputed using kNN (VIM v.6.6.2 [83]) For further information on the derivation of these variables see Additional file 3: Methods M6 and Additional file 1: Table S25 for ICD codes used. Both the EpiScores and proteins were rank inverse normalised prior to modelling. Finally, where both the EpiScore and measured protein were nominally significant ($P < 0.05$) in association with incident cardiovascular disease, nested models determined if the EpiScore improved fit by likelihood-ratio tests, (stats package in R, version 3.6.2) after adding it to a model adjusting for age, sex and the relevant (measured) protein.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03892-0.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

**Acknowledgements**

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] and is currently supported by the Wellcome Trust [216767/Z/19/Z]. Genotyping of the Generation Scotland samples was carried out by the Genetics Core Laboratory at the Edinburgh 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). The DNA methylation profiling and analysis was supported by Wellcome Investigator Award 220857/Z/20/Z and Grant 104036/Z/14/Z (PI: Prof AM McIntosh) and through funding from NARSAD (Ref: 27404; awardee: Dr DM Howard) and the Royal College of Physicians of Edinburgh (Sim Fellowship; Awardee: Prof HC Whalley).

**Peer review information**

Veronique van den Berghe and Tim Sands were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

**Authors' contributions**

J.A.R. and R.E.M. conceptualised the study design. J.A.R., A.D.C., A.Z. performed the analyses. J.B. and M.R.R. were involved software development. J.A.R., J.B., M.R.R., R.E.M. were involved in methodology. C.D., H.G., P.A., M.W., C.B., D.P., C.H. were involved in the provision of resources and study data. J.A.R., S.R.C., A.D.C., D.L.M., A.G., J.M., H.S., A.C., C.M., A.Z., M.R. were involved in data curation and validation. R.E.M., J.P., T.C.R., S.R.C. provided supervision. D.P., C.H. were responsible for acquisition of funding. All authors read and approved the final manuscript.

**Funding**

JAR is a University of Edinburgh Clinical Academic Track PhD student, supported by the Wellcome Trust (319878/Z/24/Z). ADC was supported by a Medical Research Council PhD Studentship in Precision Medicine with funding from the Medical Research Council Doctoral Training Program and the University of Edinburgh College of Medicine and Veterinary Medicine. HMS is a student on the University of Edinburgh Translational Neuroscience PhD programme funded by the Wellcome Trust (218493/Z/19/Z). CH was funded by MRC Human Genetics Unit program (QTL in Health and Disease) (grant U.MC_UU_00007/10). S.R.C. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (221890/Z/20/Z). JM and REM were supported by Alzheimer's Society project grant AS-PG-19b-010.

**Data availability**

Applications for access to Generation Scotland data can be made to access@generationscotland.org. Further details can be found at https://genscot.ed.ac.uk/for-researchers/access. All code associated with this manuscript is available open access on GitHub under the GNU General Public License version 3.0 (https://github.com/marioni-group/MSprot_Epigenetics) and Zenodo [84]. Summary statistics are available in Zenodo under a Creative Commons Attribution 4.0 International license [85].

## Declarations

### Ethics approval and consent to participate

Ethical approval for the GS cohort was received from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89) and Research Tissue Bank status was granted by the East of Scotland Research Ethics Service (REC Reference Number: 20/ES/0021). Participants provided written informed consent.

### Consent for publication

Not applicable.

### Competing interests

C.B., A.Z. and M.R. are co-founders of Eliptica Ltd. C.B.M. is a consultant and shareholder of Eliptica Ltd (London, UK). R.E.M. is a scientific advisor to the Epigenetic Clock Development Foundation and Optima Partners. D.L.M. is employed by Optima Partners Ltd. The other authors have no competing interests to declare.

### Author details

[1]Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK
[2]Institute of Science and Technology, Vienna, Austria
[3]Molecular Biology of Metabolism Laboratory, The Francis Crick Institute, London, UK
[4]Eliptica Limited, The London Cancer Hub, Cotswold Road, Sutton, London, UK
[5]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK
[6]Centre for Research on Environment, Society and Health, School of Geosciences, University of Edinburgh, Edinburgh, UK
[7]Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK
[8]Alzheimer Scotland Dementia Research Centre, Department of Psychology, University of Edinburgh, Edinburgh, UK
[9]Medical Research Council Clinical Trials Unit, University College London, London, UK
[10]Department of Psychology, The Lothian Birth Cohorts, University of Edinburgh, Edinburgh, UK
[11]Department of Biochemistry, Charité Universitätsmedizin Berlin, Berlin, Germany
[12]Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, UK
[13]Precision Proteomics Center, Swiss Institute of Allergy and Asthma Research, University of Zurich, Zurich, Switzerland
[14]Randall Centre for Cell & Molecular Biophysics, King's College London, New Hunt's House, Guy's Campus, London SE1 1UL, UK
[15]Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, Gothenburg SE-412 96, Sweden
[16]Institute of Biotechnology, Life Sciences Centre, Vilnius University, Sauletekio al. 7, Vilnius LT10257, Lithuania

Received: 28 March 2025 / Accepted: 27 November 2025

Published online: 08 December 2025

### References

1. Gadd DA, Hillary RF, Kuncheva Z, Mangelis T, Cheng Y, Dissanayake M, et al. Blood protein assessment of leading incident diseases and mortality in the UK Biobank. Nat Aging. 2024;4:939–48. https://doi.org/10.1038/s43587-024-00655-7.
2. Carrasco-Zanini J, Pietzner M, Davitte J, Surendran P, Croteau-Chonka DC, Robins C, et al. Proteomic signatures improve risk prediction for common and rare diseases. Nat Med. 2024;30:2489–98. https://doi.org/10.1038/s41591-024-03142-z.
3. Suhre K, Zaghlool S. Connecting the epigenome, metabolome and proteome for a deeper understanding of disease. J Intern Med. 2021;290:527–48. https://doi.org/10.1111/joim.13306.
4. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013;38:23–38. https://doi.org/10.1038/npp.2012.112.
5. Yousefi PD, Suderman M, Langdon R, Whitehurst O, Davey Smith G, Relton CL. DNA methylation-based predictors of health: applications and statistical considerations. Nat Rev Genet. 2022;23:369–83. https://doi.org/10.1038/s41576-022-00465-w.
6. Stevenson AJ, McCartney DL, Hillary RF, Campbell A, Morris SW, Bermingham ML, et al. Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. Clin Epigenetics. 2020;12:113. https://doi.org/10.1186/s13148-020-00903-8.
7. Wu H, Eckhardt CM, Baccarelli AA. Molecular mechanisms of environmental exposures and human disease. Nat Rev Genet Engl. 2023;24:332–44. https://doi.org/10.1038/s41576-022-00569-3.
8. Eleanor LS, Conole AJ, Stevenson S, Muñoz Maniega SE, Harris et al. Claire Green, Maria del C. Valdés Hernández,. DNA Methylation and Protein Markers of Chronic Inflammation and Their Associations With Brain and Cognitive Aging. Neurology. 2021;97:e2340–52. https://doi.org/10.1212/wnl.0000000000012997.
9. Wielscher M, Mandaviya PR, Kuehnel B, Joehanes R, Mustafa R, Robinson O, et al. DNA methylation signature of chronic low-grade inflammation and its role in cardio-respiratory diseases. Nat Commun. 2022;13:2408. https://doi.org/10.1038/s41467-022-29792-6.
10. Gadd DA, Hillary RF, McCartney DL, Zaghlool SB, Stevenson AJ, Cheng Y et al. Epigenetic scores for the circulating proteome as tools for disease prediction. Lo YMD, Ferrucci L, editors. eLife. 2022;11:e71802. https://doi.org/10.7554/eLife.71802.
11. Cheng Y, Gadd DA, Gieger C, Monterrubio-Gómez K, Zhang Y, Berta I, et al. Development and validation of DNA methylation scores in two European cohorts augment 10-year risk prediction of type 2 diabetes. Nat Aging. 2023;3:450–8. https://doi.org/10.1038/s43587-023-00391-4.

12. Ignjatovic V, Geyer PE, Palaniappan KK, Chaaban JE, Omenn GS, Baker MS, et al. Mass spectrometry-based plasma proteomics: considerations from sample collection to achieving translational data. J Proteome Res. 2019;18:4085. https://doi.org/10.1021/acs.jproteome.9b00503.
13. Kingsmore SF. Multiplexed protein measurement: technologies and applications of protein and antibody arrays. Nat Rev Drug Discov Nat Publishing Group. 2006;5:310–21. https://doi.org/10.1038/nrd2006.
14. Messner CB, Demichev V, Wang Z, Hartl J, Kustatscher G, Mülleder M, et al. Mass spectrometry-based high-throughput proteomics and its role in biomedical studies and systems biology. Proteomics. 2023;23:2200013. https://doi.org/10.1002/pmic.202200013.
15. Zhang F, Chen W, Zhu Z, Zhang Q, Nabais MF, Qi T, et al. OSCA: a tool for omic-data-based complex trait analysis. Genome Biol. 2019;20:107. https://doi.org/10.1186/s13059-019-1718-z.
16. Trejo Banos D, McCartney DL, Patxot M, Anchieri L, Battram T, Christiansen C, et al. Bayesian reassessment of the epigenetic architecture of complex traits. Nat Commun. 2020;11:2865. https://doi.org/10.1038/s41467-020-16520-1.
17. EJ Orliac, D Trejo Banos, SE Ojavee, K Lääll, R Määgi, PM Visscher, & MR Robinson. Improving GWAS discovery and genomic prediction accuracy in biobank data. Proc Natl Acad Sci USA. 2022;119(31):e2121279119. https://doi.org/10.1073/pnas.2121279119.
18. Saffari A, Silver MJ, Zavattari P, Moi L, Columbano A, Meaburn EL, et al. Estimation of a significance threshold for epigenome-wide association studies. Genet Epidemiol. 2018;42:20–33. https://doi.org/10.1002/gepi.22086.
19. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res Engl. 2021;49:D605–12. https://doi.org/10.1093/nar/gkaa1074.
20. Battram T, Yousefi P, Crawford G, Prince C, Sheikhali Babaei M, Sharp G, et al. The EWAS catalog: a database of epigenome-wide association studies. Wellcome Open Res. 2022;7:41. https://doi.org/10.12688/wellcomeopenres.17598.2.
21. Gadd DA, Hillary RF, McCartney DL, Shi L, Stolicyn A, Robertson NA, et al. Integrated methylome and phenome study of the circulating proteome reveals markers pertinent to brain health. Nat Commun. 2022;13:4670. https://doi.org/10.1038/s41467-022-32319-8.
22. Walker RM, McCartney DL, Carr K, Barber M, Shen X, Campbell A, et al. Data resource profile: whole-blood DNA methylation resource in Generation Scotland (MeGS). Int J Epidemiol. 2025;54:dyaf091. https://doi.org/10.1093/ije/dyaf091.
23. Hillary RF, McCartney DL, Smith HM, Bernabeu E, Gadd DA, Chybowska AD, et al. Blood-based epigenome-wide analyses of 19 common disease states: a longitudinal, population-based linked cohort study of 18,413 Scottish individuals. PLoS Med. 2023;20:e1004247. https://doi.org/10.1371/journal.pmed.1004247.
24. Hedman ÅK, Mendelson MM, Marioni RE, Gustafsson S, Joehanes R, Irvin MR, et al. Epigenetic patterns in blood associated with lipid traits predict incident coronary heart disease events and are enriched for results from genome-wide association studies. Circ Cardiovasc Genet. 2017;10:e001487. https://doi.org/10.1161/CIRCGENETICS.116.001487.
25. Cardona A, Day FR, Perry JRB, Loh M, Chu AY, Lehne B, et al. Epigenome-wide association study of incident type 2 diabetes in a British population: EPIC-Norfolk study. Diabetes. 2019;68:2315–26. https://doi.org/10.2337/db18-0290.
26. Zaghlool SB, Kühnel B, Elhadad MA, Kader S, Halama A, Thareja G, et al. Epigenetics meets proteomics in an epigenome-wide association study with circulating blood plasma protein traits. Nat Commun. 2020;11:15. https://doi.org/10.1038/s41467-019-13831-w.
27. Smith HM, Ng HK, Moodie JE, Gadd DA, McCartney DL, Bernabeu E, et al. DNA methylation-based predictors of metabolic traits in Scottish and Singaporean cohorts. Am J Hum Genet. 2025;112:106–15. https://doi.org/10.1016/j.ajhg.2024.11.012.
28. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017;541:81–6. https://doi.org/10.1038/nature20784.
29. Nuotio M-L, Pervjakova N, Joensuu A, Karhunen V, Hiekkalinna T, Milani L, et al. An epigenome-wide association study of metabolic syndrome and its components. Sci Rep. 2020;10:20567. https://doi.org/10.1038/s41598-020-77506-z.
30. Schmitz G, Langmann T, Heimerl S. Role of ABCG1 and other ABCG family members in lipid metabolism. J Lipid Res. 2001;42:1513–20. https://doi.org/10.1016/S0022-2275(20)32205-7.
31. Aslibekyan S, Claas SA. Methylation in CPT1A, Lipoproteins, and epigenetics. In: Patel V, Preedy V, editors. Handb Nutr diet epigenetics [Internet]. Cham: Springer International Publishing; 2017. pp. 1–17. https://doi.org/10.1007/978-3-319-31143-2_108-1.
32. Hillary RF, Ng HK, McCartney DL, Elliott HR, Walker RM, Campbell A, et al. Blood-based epigenome-wide analyses of chronic low-grade inflammation across diverse population cohorts. Cell Genom. 2024;4:100544. https://doi.org/10.1016/j.xgen.2024.100544.
33. Choi E-H, Park S-J. A key protein in the cellular stress response pathway and a potential therapeutic target. Exp Mol Med. 2023;55:1348–56. https://doi.org/10.1038/s12276-023-01019-8.
34. Rohm TV, Meier DT, Olefsky JM, Donath MY. Inflammation in obesity, diabetes, and related disorders. Immun Elsevier. 2022;55:31–55. https://doi.org/10.1016/j.immuni.2021.12.013.
35. Kobayashi KS, van den Elsen PJ. NLRC5: a key regulator of MHC class I-dependent immune responses. Nat Rev Immunol. 2012;12:813–20. https://doi.org/10.1038/nri3339.
36. Mugoni V, Postel R, Catanzaro V, De Luca E, Turco E, Digilio G, et al. Ubiad1 is an antioxidant enzyme that regulates eNOS activity by CoQ10 synthesis. Cell. 2013;152:504–18. https://doi.org/10.1016/j.cell.2013.01.013.
37. McCartney DL, Zhang F, Hillary RF, Zhang Q, Stevenson AJ, Walker RM, et al. An epigenome-wide association study of sex-specific chronological ageing. Genome Med. 2019;12:1. https://doi.org/10.1186/s13073-019-0693-z.
38. Preissner KT, Seiffert D. Role of vitronectin and its receptors in haemostasis and vascular remodeling. Thromb Res. 1998;89:1–21. https://doi.org/10.1016/s0049-3848(97)00298-3.
39. Ekmekci H, Sonmez H, Ekmekci OB, Ozturk Z, Domanic N, Kokoglu E. Plasma vitronectin levels in patients with coronary atherosclerosis are increased and correlate with extent of disease. J Thromb Thrombolysis. 2002;14:221–5. https://doi.org/10.1023/A:1025000810466.
40. Aslan S, Ikitimur B, Cakmak HA, Ozcan S, Yuksel H. Vitronectin levels and coronary artery disease severity in acute coronary syndromes. Eur Heart J. 2013;34:P4053. https://doi.org/10.1093/eurheartj/eht309.P4053.
41. Derer W, Barnathan ES, Safak E, Agarwal P, Heidecke H, Möckel M, et al. Vitronectin concentrations predict risk in patients undergoing coronary stenting. Circ Cardiovasc Interv. 2009;2:14–9. https://doi.org/10.1161/CIRCINTERVENTIONS.108.795799.

42. Shin M, Park SH, Mun S, Lee J, Kang H-G. Biomarker discovery of acute coronary syndrome using proteomic approach. Molecules. 2021. https://doi.org/10.3390/molecules26041136.
43. Asada M, Oishi E, Sakata S, Hata J, Yoshida D, Honda T, et al. Serum lipopolysaccharide-binding protein levels and the incidence of cardiovascular disease in a general Japanese population: the Hisayama study. J Am Heart Assoc. 2019;8:e013628. https://doi.org/10.1161/JAHA.119.013628.
44. Tollefsen DM. Heparin cofactor II modulates the response to vascular injury. Arterioscler Thromb Vasc Biol. 2007;27:454–60. https://doi.org/10.1161/01.ATV.0000256471.22437.88.
45. Huang P-H, Leu H-B, Chen J-W, Wu T-C, Lu T-M, Yu-An Ding P, et al. Decreased heparin cofactor II activity is associated with impaired endothelial function determined by brachial ultrasonography and predicts cardiovascular events. Int J Cardiol. 2007;114:152–8. https://doi.org/10.1016/j.ijcard.2005.12.009.
46. Zhang Q, Wen X-H, Tang S-L, Zhao Z-W, Tang C-K. Role and therapeutic potential of gelsolin in atherosclerosis. J Mol Cell Cardiol. 2023;178:59–67. https://doi.org/10.1016/j.yjmcc.2023.03.012.
47. Zheng SC, Beck S, Jaffe AE, Koestler DC, Hansen KD, Houseman AE, et al. Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses. Nat Methods. 2017;14:216–7. https://doi.org/10.1038/nmeth.4187.
48. Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects*. Mol Cell Proteomics. 2002;1:845–67. https://doi.org/10.1074/mcp.R200007-MCP200.
49. Georgila K, Vyrla D, Drakos E. Apolipoprotein A-I (ApoA-I), immunity, inflammation and cancer. Cancers. 2019;11:1097. https://doi.org/10.3390/cancers11081097.
50. Angelov A, Connelly PJ, Delles C, Kararigas G. Sex-biased and sex hormone-dependent regulation of apolipoprotein A1. Curr Opin Physiol. 2023;33:100654. https://doi.org/10.1016/j.cophys.2023.100654.
51. Frondelius K, Borg M, Ericson U, Borné Y, Melander O, Sonestedt E. Lifestyle and dietary determinants of serum apolipoprotein A1 and apolipoprotein B concentrations: cross-sectional analyses within a Swedish cohort of 24,984 individuals. Nutrients. 2017;9:211. https://doi.org/10.3390/nu9030211.
52. Cabana VG, Siegel JN, Sabesin SM. Effects of the acute phase response on the concentration and density distribution of plasma lipids and apolipoproteins. J Lipid Res. 1989;30:39–49. https://doi.org/10.1016/S0022-2275(20)38390-5.
53. Faaborg-Andersen CC, Liu C, Subramaniyam V, Desai SR, Sun YV, Wilson PWF, et al. U-shaped relationship between apolipoprotein A1 levels and mortality risk in men and women. Eur J Prev Cardiol. 2023;30:293–304. https://doi.org/10.1093/eurjpc/zwac263.
54. Ghaemi F, Rabizadeh S, Yadegar A, Mohammadi F, Asadigandomani H, Bafrani MA, et al. ApoA1/HDL-C ratio as a predictor for coronary artery disease in patients with type 2 diabetes: a matched case-control study. BMC Cardiovasc Disord. 2024;24:317. https://doi.org/10.1186/s12872-024-03986-w.
55. Cui M, Cheng C, Zhang L. High-throughput proteomics: a methodological mini-review. Lab Invest. 2022;102:1170–81. https://doi.org/10.1038/s41374-022-00830-7.
56. da Costa JP, Santos PSM, Vitorino R, Rocha-Santos T, Duarte AC. How low can you go? A current perspective on low-abundance proteomics. TrAC Trends Anal Chem. 2017;93:171–82. https://doi.org/10.1016/j.trac.2017.05.014.
57. Carreras-Gallo N, Chen Q, Balagué-Dobón L, Aparicio A, Giosan IM, Dargham R et al. Leveraging DNA methylation to create epigenetic biomarker proxies that inform clinical care: A new framework for precision medicine. MedRxiv. 2024;2024.12.06.24318612. https://doi.org/10.1101/2024.12.06.24318612.
58. Chybowska AD, Bernabeu E, Yousefi P, Suderman M, Hillary RF, Clark R, et al. A blood- and brain-based EWAS of smoking. Nat Commun. 2025;16:3210. https://doi.org/10.1038/s41467-025-58357-6.
59. Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM, et al. Cohort profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. Int J Epidemiol. 2012;42:689–700. https://doi.org/10.1093/ije/dys084.
60. Messner CB, Demichev V, Bloomfield N, Yu JSL, White M, Kreidl M, et al. Ultra-fast proteomics with scanning SWATH. Nat Biotechnol. 2021;39:846–54. https://doi.org/10.1038/s41587-021-00860-4.
61. Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat Methods. 2020;17:41–4. https://doi.org/10.1038/s41592-019-0638-x.
62. Vernardis SI, Demichev V, Lemke O, Grüning N-M, Messner C, White M, et al. The impact of acute nutritional interventions on the plasma proteome. J Clin Endocrinol Metab. 2023;108:2087–98. https://doi.org/10.1210/clinem/dgad031.
63. Bruderer R, Muntel J, Müller S, Bernhardt OM, Gandhi T, Cominetti O, et al. Analysis of 1508 plasma samples by capillary-flow data-independent acquisition profiles proteomics of weight loss and maintenance. Mol Cell Proteomics. 2019;18:1242–54. https://doi.org/10.1074/mcp.RA118.001288.
64. Rusilowicz M, Dickinson M, Charlton A, O'Keefe S, Wilson J. A batch correction method for liquid chromatography–mass spectrometry data that does not depend on quality control samples. Metabolomics. 2016;12:56. https://doi.org/10.1007/s11306-016-0972-2.
65. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47. https://doi.org/10.1093/nar/gkv007.
66. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51:D523-31. https://doi.org/10.1093/nar/gkac1052.
67. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the illumina methylation-epic BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;17:208. https://doi.org/10.1186/s13059-016-1066-1.
68. van Iterson M, van Zwet EW, Heijmans BT. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. Genome Biol. 2017;18:19. https://doi.org/10.1186/s13059-016-1131-9.
69. Therneau TM, coxme. Mixed Effects Cox Models. R package version 2.2–22. 2024. https://CRAN.R-project.org/package=coxme.
70. Campagna MP, Xavier A, Lechner-Scott J, Maltby V, Scott RJ, Butzkueven H, et al. Epigenome-wide association studies: current knowledge, strategies and recommendations. Clin Epigenetics. 2021;13:214. https://doi.org/10.1186/s13148-021-01200-8.
71. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics. 2014;30:1431–9. https://doi.org/10.1093/bioinformatics/btu029.

72. Rockel T, missMethods. Methods for Missing Data. 2022; https://github.com/torockel/missMethods.
73. Hillary RF, Trejo-Banos D, Kousathanas A, McCartney DL, Harris SE, Stevenson AJ, et al. Multi-method genome- and epigenome-wide studies of inflammatory protein levels in healthy older adults. Genome Med. 2020;12:60. https://doi.org/10.1186/s13073-020-00754-1.
74. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9. https://doi.org/10.1093/bioinformatics/btu049.
75. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomart. Nat Protoc. 2009;4:1184–91. https://doi.org/10.1038/nprot.2009.97.
76. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. Nucleic Acids Res. 2023;51:D933–41. https://doi.org/10.1093/nar/gkac958.
77. Nagy R, Boutin TS, Marten J, Huffman JE, Kerr SM, Campbell A, et al. Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 generation Scotland participants. Genome Med. 2017;9:23. https://doi.org/10.1186/s13073-017-0414-4.
78. Noguera-Castells A, García-Prieto CA, Álvarez-Errico D, Esteller M. Validation of the new EPIC DNA methylation microarray (900K EPIC v2) for high-throughput profiling of the human DNA methylome. Epigenetics. 2023;18:2185742. https://doi.org/10.1080/15592294.2023.2185742.
79. Cox DR. Regression models and life-tables. J R Stat Soc Ser B Stat Methodol. 1972;34:187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.
80. Kuan V, Denaxas S, Gonzalez-Izquierdo A, Direk K, Bhatti O, Husain S, et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the english National health service. Lancet Digit Health Elsevier. 2019;1:e63–77. https://doi.org/10.1016/S2589-7500(19)30012-3.
81. Welsh P, Preiss D, Hayward C, Shah ASV, McAllister D, Briggs A, et al. Cardiac troponin T and troponin I in the general population. Circulation. 2019;139:2754–64. https://doi.org/10.1161/CIRCULATIONAHA.118.038529.
82. Therneau TA, Package for. Survival Analysis in R. R package version 3.5-8. 2024. https://CRAN.R-project.org/package=survival.
83. Kowarik A, Templ M. Imputation with the R package VIM. J Stat Softw. 2016;74:1–16. https://doi.org/10.18637/jss.v074.i07.
84. Robertson J, Bajzik J, Vernardis S, Chybowska A, McCartney D, Grauslys A et al. Methylome-wide association studies and epigenetic biomarker development for 133 mass spectrometry-assessed Circulating proteins in 14,761 generation Scotland participants. GitHub; 2025. https://doi.org/10.5281/zenodo.17589713.
85. Robertson J, Bajzík J, Vernardis S, Chybowska A, Daniel M, Grauslys A et al. Mass spectrometry proteins: EWAS and episcores in generation Scotland. Zenodo; 2025. https://doi.org/10.5281/zenodo.16924748.

## Publisher's Note