

Methylome-wide studies of six metabolic traits

Hannah M. Smith¹, Hong Kiat Ng², Joanna E. Moodie³, Danni A. Gadd¹, Daniel L. McCartney¹, Elena Bernabeu¹, Archie Campbell¹, Paul Redmond³, Adele Taylor³, Danielle Page³, Janie Corley³, Sarah E. Harris³, Darwin Tay², Ian J. Deary³, Kathryn L. Evans¹, Matthew R. Robinson⁴, John C. Chambers^{2,5}, Marie Loh^{2,5,6}, Simon R. Cox³, Riccardo E. Marioni^{1*}, Robert F. Hillary^{1*}

¹Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

²Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

³Lothian Birth Cohorts, Department of Psychology, University of Edinburgh, Edinburgh, UK

⁴Institute of Science and Technology Austria, Am Campus 1, 3400 Klosterneuburg, Austria

⁵Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK

⁶Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), Singapore

*equal contribution

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

Word count: 3,534

Abstract

Exploring the molecular correlates of metabolic health measures may identify the shared and unique biological processes and pathways that they track. Here, we performed epigenome-wide association studies (EWASs) of six metabolic traits: body mass index (BMI), body fat percentage, waist-hip ratio (WHR), and blood-based measures of glucose, high-density lipoprotein (HDL) cholesterol, and total cholesterol. We considered blood-based DNA methylation (DNAm) from >750,000 CpG sites in over 17,000 volunteers from the Generation Scotland (GS) cohort. Linear regression analyses identified between 304 and 11,815 significant CpGs per trait at $P < 3.6 \times 10^{-8}$, with 37 significant CpG sites across all six traits. Further, we performed a Bayesian EWAS that jointly models all CpGs simultaneously and conditionally on each other, as opposed to the marginal linear regression analyses. This identified between 3 and 27 CpGs with a posterior inclusion probability ≥ 0.95 across the six traits. Next, we used elastic net penalised regression to train epigenetic scores (EpiScores) of each trait in GS, which were then tested in the Lothian Birth Cohort 1936 (LBC1936; European ancestry) and Health for Life in Singapore (HELIOS; Indian-, Malay- and Chinese-ancestries). A maximum of 27.1% of the variance in BMI was explained by the BMI EpiScore in the subset of Malay-ancestry Singaporeans. Four metabolic EpiScores were associated with general cognitive function in LBC1936 in models adjusted for vascular risk factors (Standardised β_{range} : 0.08 – 0.12, $P_{\text{FDR}} < 0.05$). EpiScores of metabolic health are applicable across ancestries and can reflect differences in brain health.

Introduction

Measures of adiposity and lipids are central to profiling metabolic health. There are several clinical measures of metabolic health, which include body mass index (BMI), body fat percentage, waist-hip ratio (WHR), blood glucose levels, high-density lipoprotein (HDL) cholesterol, and total cholesterol. These traits have routinely been linked to health-related risks including cardiovascular disease (1-3), myocardial infarction (4), and stroke (2, 3, 5). Multiple associations between metabolic traits and cognitive function and rate of cognitive decline have also been observed (6-12). BMI is a widely assessed indicator of metabolic health but is limited by its inability to directly track the amount or distribution of fat in the body (13, 14). BMI has previously shown low specificity in identifying individuals with excess body fat (15). Considering multiple measures that track different aspects of adiposity (and related traits) may provide a more complete assessment of metabolic health. Furthermore, exploring the molecular correlates of these metabolic indices may help to inform the shared and unique biological processes and pathways that they are associated with.

The epigenetic modification DNA methylation (DNAm) is dynamic, tissue/cell-type specific, and can be affected by genetic and environmental factors. Epigenome-wide association studies (EWASs) have detailed associations between individual blood-based DNAm loci (CpG sites) and metabolic traits including BMI, WHR, HDL cholesterol, and total cholesterol (16-32). In our previous work, penalised regression models have been applied to DNAm data to develop molecular predictors for a multitude of complex traits. These epigenetic scores, or EpiScores, may augment associations with health outcomes when combined with their measured phenotypic counterparts (33-35). For example, an EpiScore for BMI increased the amount of variance in metabolic health outcomes accounted for by measured BMI alone by an

average of 3% (36). An EpiScore for WHR was also associated with all-cause mortality in the same population of healthy older adults after adjusting for measured WHR (33). Here, we modelled EWASs with both linear regression and Bayesian penalised regression on six metabolic traits in the Generation Scotland (GS) study ($N > 17,000$). In the former approach, we obtained marginal estimates for each CpG, which do not take into account correlations across CpGs. By contrast, the Bayesian penalised regression estimated CpG effects jointly so that the effect of each CpG was conditional on all other loci. We compared findings from the individual EWASs to determine whether the six traits showed unique or common methylomic signatures. We then trained EpiScores for the six metabolic traits in GS ($N > 17,000$) and projected them into two independent test cohorts – the Lothian Birth Cohort 1936 (LBC1936) and the Health for Life in Singapore (HELIOS) cohort. Finally, we tested metabolic trait EpiScore associations with general cognitive function level and change in LBC1936 ($N = 861$). Associations identified between EpiScores for metabolic traits and cognitive phenotypes could offer new opportunities to examine the relevance of metabolic health indicators to ageing, and cognitive and neurological health outcomes. A visual summary of the study is shown in **Figure 1**.

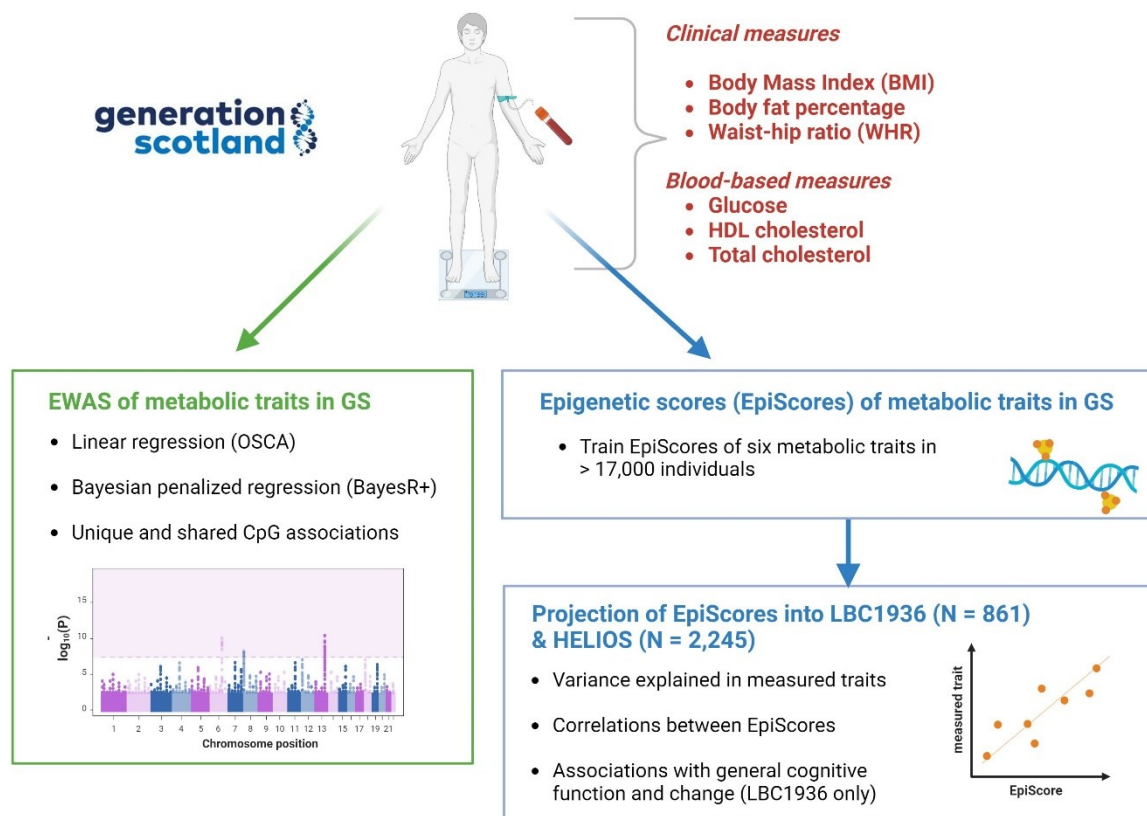


Figure 1: Summary of metabolic trait study. This figure provides an overview of the analysis performed in this study. Created with BioRender.com.

Methods

Generation Scotland Cohort

The Generation Scotland (GS) Cohort has been described in detail previously (37). Briefly, it is a Scotland-wide, family-based study of health. In the current study, 18,411 individuals had DNA methylation profiled on the Illumina EPIC array from blood samples taken at the study baseline between 2006 and 2011. Quality control (QC) details can be found in **Supplementary Methods**. 59% of the cohort was female and the mean age at baseline was 47.5 years (SD: 14.9). Six metabolic measures from GS were utilised in this study: body mass index (BMI, kg/m^2), body fat percentage, waist-hip ratio (WHR), glucose (mmol/L), serum HDL cholesterol (mmol/L), and serum total cholesterol (mmol/L) (**Table 1, Supplementary Methods**).

The Lothian Birth Cohort 1936

The Lothian Birth Cohort 1936 (LBC1936) is a longitudinal study of ageing (38, 39). The study consists of individuals born in 1936, most of whom sat a general cognitive ability test at a mean age of 11 years in Scotland. Individuals living in the Lothian area were recruited to the LBC1936 study at around age 70 (baseline N=1,091). The volunteers undertook triennial testing across five waves of follow-up (ages ~70, 73, 76, 79, and 82). Of those with blood-based DNA methylation data (profiled on the Illumina 450k array) at wave 1, the mean age was 69.6 years (SD: 0.83) with 49.4% females. QC and pre-processing for the DNA methylation in the LBC1936 can be found in **Supplementary Methods**. Three metabolic measures were utilised in this study: BMI (kg/m²), serum HDL cholesterol (mmol/L), and serum total cholesterol (mmol/L) (**Table 1, Supplementary Methods**). Thirteen cognitive tests were assessed longitudinally (details in **Supplementary Methods**).

The Health for Life in Singapore cohort

The Health for Life in Singapore (HELIOS) study is a single-centre, multi-ancestry cohort of approximately 10,000 individuals residing in Singapore. A subset of the cohort in which Illumina EPIC DNA methylation data have been profiled has a mean age of 54.3 (SD: 11.7) and 61.2% of the cohort was female. The subset is made up of three self-reported ancestry groups: Chinese and other East Asian (Chinese) (N = 1,778), Malay and other South-East-Asian (Malay) (N = 242), and South Asian (Indian and other countries from the Indian subcontinent) (N = 225). QC and pre-processing of DNA methylation in HELIOS can be found in the **Supplementary Methods**. Five metabolic measures were utilised in this study: BMI (kg/m²), body fat percentage, WHR, serum HDL cholesterol (mmol/L) and serum total cholesterol (mmol/L) (**Table 1, Supplementary Methods**).

Epigenome-wide association studies of six metabolic traits in GS

Linear regression models tested for associations between 752,722 CpG sites and each of the six metabolic traits in GS using the fast linear method in the OmicS-data-based Complex trait Analysis (OSCA) software (40). To facilitate less computationally expensive analyses, phenotypes were regressed on age, age², sex and family structure (to account for relatedness in GS) (41)) using linear mixed-effects models (lmeKin function from the coxme package (version: 2.2.18.1) in R) (42). Family structure was modelled with a kinship matrix constructed using the R package kinship2 (version: 1.9.6). CpG M-values were pre-corrected for age, sex and experimental batch ($N = 121$ batches) in linear regression models using the lm function in R. Residuals from the regression models for each outcome trait and CpG were taken forward for the EWASs. An epigenetic smoking score, EpiSmokEr, derived using the SSc method which adds up methylation levels of 187 CpG sites found to be significantly associated with smoking in a study by Zeilinger *et al* (43, 44) and Houseman-estimated white blood cell proportions (45) were included as fixed-effect covariates in the OSCA analysis. Finally, the first 20 methylation principal components (PCs) were included as covariates to account for potentially unmeasured confounders. Descriptive statistics can be found in **Supplementary Table 1**. A significance level of $P < 3.6 \times 10^{-8}$ was set to detect significantly associated CpGs as suggested by Saffari *et al* in a study investigating significance thresholds in EWAS using a simulation approach (46). Mapping of CpG sites to genes was performed using Illumina annotation files. Principal component analyses (PCA) were performed on the significantly associated CpG sites from each metabolic trait EWAS. The number of approximate independent signals was denoted as the cumulative number of principal components that accounted for at least 80% of the variance among all significantly

associated probes. PCA was performed using the scikit-learn package in Python (2.7.17) (47).

Measure	N	Mean	SD	Range
Generation Scotland				
Age (years)	18,411	47.5	14.9	17.1 to 98.5
BMI (kg/m ²)	17,304	26.5	4.7	17 to 49
Body fat (%)	17,304	29.8	9.1	8 to 50
WHR	17,304	0.9	0.1	0.4 to 1.4
Glucose (mmol/L)	17,908	4.7	0.6	1.3 to 9.2
HDL cholesterol (mmol/L)	18,225	1.5	0.4	0.4 to 3.1
Total cholesterol (mmol/L)	18,270	5.1	1.1	0.9 to 9.3
Lothian Birth Cohort 1936				
Age (years)	861	69.6	0.8	67.7 to 70.4
BMI (kg/m ²)	860	27.8	4.3	16 to 47.3
HDL cholesterol (mmol/L)	779	1.5	0.4	0.5 to 3.8
Total cholesterol (mmol/L)	851	5.4	1.2	2.7 to 10.8
Health for Life in Singapore				
Age (years)	2,245	54.3	11.7	30.2 to 85.4
BMI (kg/m ²)	2,226	24.1	1.2	14.2 to 43.7
Body fat (%)	2,063	38.2	7.2	17.6 to 63.1
WHR	2,233	0.9	0.1	0.67 to 1.1
HDL cholesterol (mmol/L)	2,227	1.5	0.4	0.7 to 3

Total cholesterol (mmol/L)	2,223	5.3	1	2.4 to 8.6
----------------------------	-------	-----	---	------------

Table 1: Cohort demographics for Generation Scotland, the Lothian Birth Cohort 1936 and the Health for Life in Singapore study. Table 1 shows the demographics of the data included in this study including N, mean, range and standard deviation for each variable after outlier removal. SD = standard deviation, BMI = body mass index, WHR = waist-hip ratio, HDL = high-density lipoprotein.

Gene ontology enrichment analysis

We tested whether common CpGs identified across all six marginal linear regression EWAS models were over-represented among gene ontology (GO) terms using the gometh function from the missMethyl R package version 1.34 (48). The probability of significant differential methylation due to the number of probes per gene was taken into consideration. Statistically significant results were defined as having $P_{FDR} < 0.05$.

Bayesian EWAS

Probe-by-probe (marginal) linear regression models fail to consider the correlation structure that exists across the methylome. Therefore, we considered Bayesian penalised regression, conducted using BayesR+ (49), as a secondary analysis. This method estimates single marker or probe effects whilst controlling for all other probes as well as being able to control for known and unknown confounding variables. This method also estimates the amount of phenotypic variation attributed to genome-wide DNA methylation. We applied the same covariate and phenotype preparation strategy as in the linear regression models. Significant CpGs were defined as sites with a posterior inclusion probability (PIP) ≥ 0.95 . Details on the methods used for the Bayesian strategy can be found in the **Supplementary Methods**.

Replication of previous literature

The EWAS catalogue (16) was used to determine if the overlapping CpGs that were found to be associated with all six metabolic traits in the linear regression EWASs have previously been identified in other studies. The EWAS catalogue was filtered to whole blood samples, CpG-metabolic trait associations with $P < 3.6 \times 10^{-8}$ (in line with our study and consistent with Saffari, *et al* (46)) and study sample $N > 1,000$ participants. The search terms used to identify traits from the EWAS catalogue can be found in **Supplementary Table 2**. The EWAS catalogue was not filtered for studies that may contain GS data.

Generation and projection of DNA methylation-based proxies of six metabolic traits

Penalised regression models were trained in GS to generate epigenetic scores (EpiScores) of each of the six metabolic traits using the R package biglasso (version 1.5.2). Each trait was modelled as the response variable (using the same phenotype files from the EWASs) and 395,380 CpGs (the 450K methylation array subset that was present in GS) were used as predictors. Cross-validation was carried out ($n_{\text{folds}} = 20$) and an elastic net (elnet) penalty was set ($\alpha = 0.5$). CpG sites with a non-zero coefficient were retained and used to derive EpiScores in LBC1936 ($n = 861$). This was followed by further testing in the HELIOS cohort ($n = 2,245$). Missing CpGs were mean imputed in LBC1936 and HELIOS. Predictors obtained from the Bayesian penalised regression models were also projected into LBC1936 and HELIOS using the mean posterior effect sizes as weights for the scores. The variance explained (incremental R^2) in each metabolic trait by their corresponding EpiScore over and above age and sex in linear regression models was then calculated. In HELIOS, the variance explained was calculated in the full cohort and ancestry subgroups. In HELIOS full cohort models, additional adjustments for ancestry were included.

EpiScore associations with general cognitive function and change in LBC1936

A latent intercept and age-related slope for general cognitive function were generated in LBC1936 using a structural equation modelling (SEM) framework with the R package Lavaan (version 0.6.12) (50). Measured traits and EpiScores were regressed on intercepts and slopes in separate linear models. Full details are provided in **Supplementary Methods** and **Supplementary Tables 3-6**.

Results

Epigenome Wide Association Studies (EWASs) of six metabolic traits

Correlations between metabolic traits in GS ranged between -0.36 (WHR and HDL cholesterol) and 0.6 (BMI and body fat percentage), and are shown in **Supplementary Figure 1**. Marginal linear regression EWASs of six metabolic traits were performed in GS. The number of CpG sites significantly associated ($P < 3.6 \times 10^{-8}$) with each of the traits are summarised in **Table 2**. This ranged between 304 for glucose to 11,815 for BMI. Manhattan plots can be observed in **Supplementary Figure 2** and the top 1,000 significantly associated CpGs with each trait are listed in **Supplementary Table 7**. Full summary statistic output will be available upon publication.

The large number of significant associations observed in our models may reflect correlation structures among CpG sites (Quantile-Quantile plots and inflation factors – which ranged between 1.18 and 2.48 – can be observed in **Supplementary Figure 3** and **Supplementary Table 8**). Therefore, we performed PCA for each trait to determine the approximate number of independent features present among CpG sites that surpassed the epigenome-wide significance threshold ($P < 3.6 \times 10^{-8}$). We identified between 80 and 1,302 (for glucose and BMI, respectively) principal

components or ‘independent features’ that accounted for $\geq 80\%$ of the variance in the underlying CpG sites (**Table 2**).

Next, we performed Bayesian penalised regression, which jointly models all CpGs and accounts for genome-wide correlation patterns. **Table 2** shows the number of high-confidence associations ($PIP \geq 0.95$), which ranged between 3 (glucose) and 27 associations (BMI) (**Supplementary Table 9**). Using the Bayesian method, we obtained estimates for the variance captured by genome-wide DNA methylation that ranged between 24% for WHR and 53% for BMI (**Supplementary Table 10**).

Trait	Number of CpGs in marginal linear regression EWASs at $P < 3.6 \times 10^{-8}$	Number of PCs that explain $\geq 80\%$ of variance in the significant CpGs from linear regression models	Number of CpGs in Bayesian EWASs at $PIP \geq 0.95$
BMI	11,815	1,302	27
WHR	4,334	687	12
Body fat percentage	8,468	1,189	18
Glucose	304	80	3
HDL cholesterol	7,623	1,085	20
Total cholesterol	1,722	328	19

Table 2: The number of significantly associated CpGs with each metabolic trait in Generation Scotland.

The table shows the number of significantly associated CpGs with each metabolic trait using marginal linear regression and Bayesian penalised regression. The table also shows the number of principal components that account for $\geq 80\%$ of the variance of the significant CpGs from the linear regression analyses for each metabolic trait. BMI = body mass index, WHR = waist-hip ratio, HDL = high-density lipoprotein, PCs = principal components.

37 CpG sites were significant ($P < 3.6 \times 10^{-8}$) across all six metabolic traits in the marginal linear regression models (**Supplementary Table 11, Supplementary Figure 4**). In the Bayesian models, a single CpG site, “cg06500161” (mapped to the

ABCG1 gene), was associated with BMI, body fat percentage, HDL cholesterol, total cholesterol, and WHR (PIP ≥ 0.95 , **Supplementary Table 9**).

14 of the 37 common CpGs from the linear models had been previously associated with metabolic traits in studies using whole blood samples at $P < 3.6 \times 10^{-8}$ and study $N > 1000$ reported in the EWAS catalogue (**Supplementary Table 11**). Of the 37 CpGs associated with all traits in the linear models, four mapped to the *CPT1A* gene, four mapped to the *ABCG1* gene, and three mapped to the PHGDH gene. Seven of the overlapping CpGs did not map to any genes. The remaining 19 CpGs mapped to unique genes giving a total of 22 unique genes containing the overlapping CpGs. Gene ontology (GO) enrichment analysis of the 37 common CpGs was performed. Eleven GO terms were found to be enriched, including cholesterol biosynthetic process and regulation of lipid storage. The full list of enriched GO terms identified can be found in **Supplementary Table 12**.

Epigenetic Scores (EpiScores) of metabolic traits tested in the LBC1936 and HELIOS

EpiScores for each of the six metabolic traits were trained in GS using elastic net (elnet) penalised regression and projected into the LBC1936 and HELIOS cohorts. We explored how much additional variance could be accounted for in each metabolic trait by the corresponding EpiScore over and above linear regression models adjusting for age and sex. In the LBC1936, EpiScores accounted for 3.2% of the variance for total cholesterol, 18.5% for HDL cholesterol, and 14.4% of the variance in BMI. In HELIOS full cohort analysis, the incremental R^2 estimates ranged between 7.1% (for total cholesterol) to 20.8% (for BMI). However, there was variability within the ancestry-specific subsets of HELIOS. Most notably, the body fat percentage EpiScore

accounted for 9.2% and 9.5% in the Chinese and Malay subgroups but only 3.1% in the Indian subgroup (**Figure 2, Supplementary Table 13**). In LBC1936 and HELIOS, the correlations between all six EpiScores are shown in **Supplementary Figure 5**. Correlations between measured traits ranged from -0.3 – 0.38 for LBC1936, and -0.46 - 0.47 for HELIOS (**Supplementary Figure 6**). Correlations between measured traits and EpiScores ranged between -0.41 – 0.5 in LBC1936 and -0.66 – 0.92 in HELIOS (**Supplementary Figure 7**).

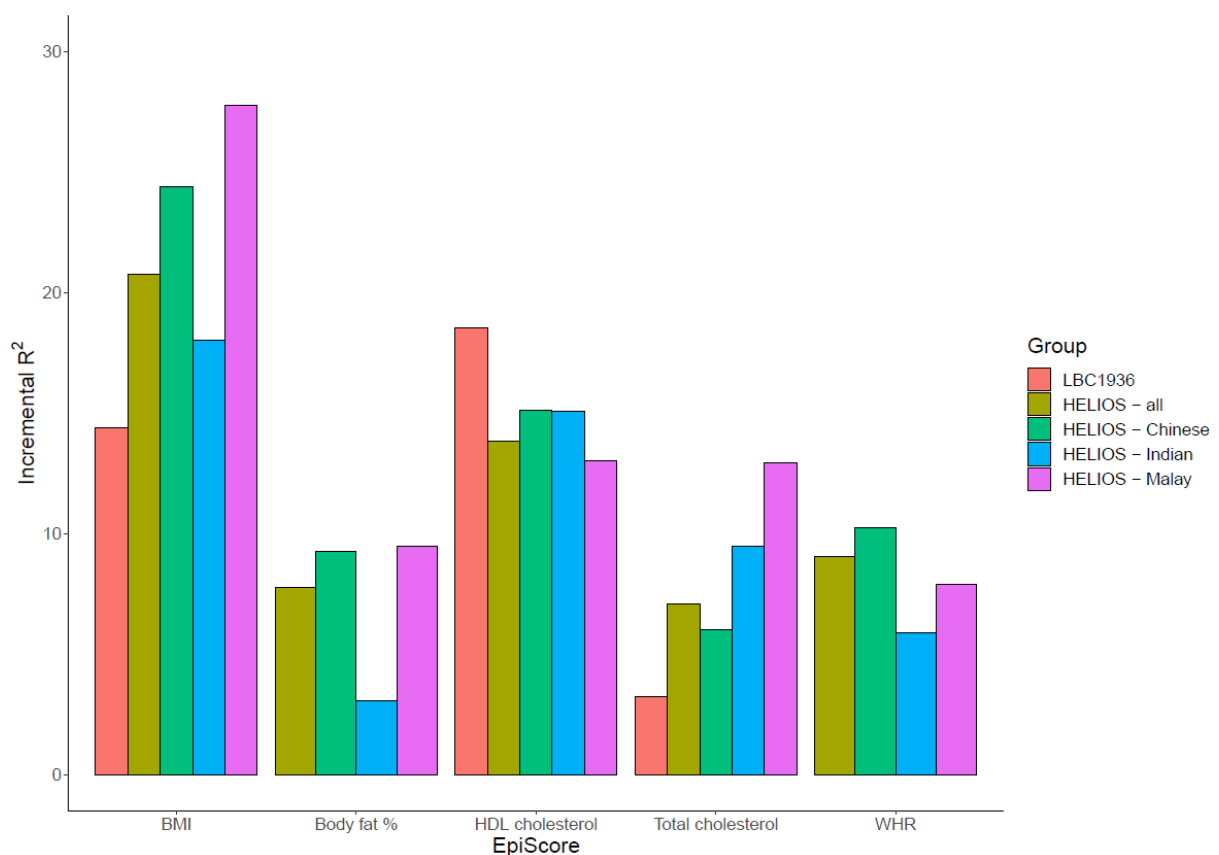


Figure 2: The variance explained in measured metabolic traits by elnet EpiScores in the Lothian Birth Cohort 1936 (LBC1936) and the Health for Life in Singapore (HELIOS) study. Additional variance (incremental R²) accounted for in each metabolic trait by their corresponding elnet EpiScores over and above age and sex-adjusted (and ancestry in HELIOS full cohort) linear regression models in LBC1936 and HELIOS. Measured

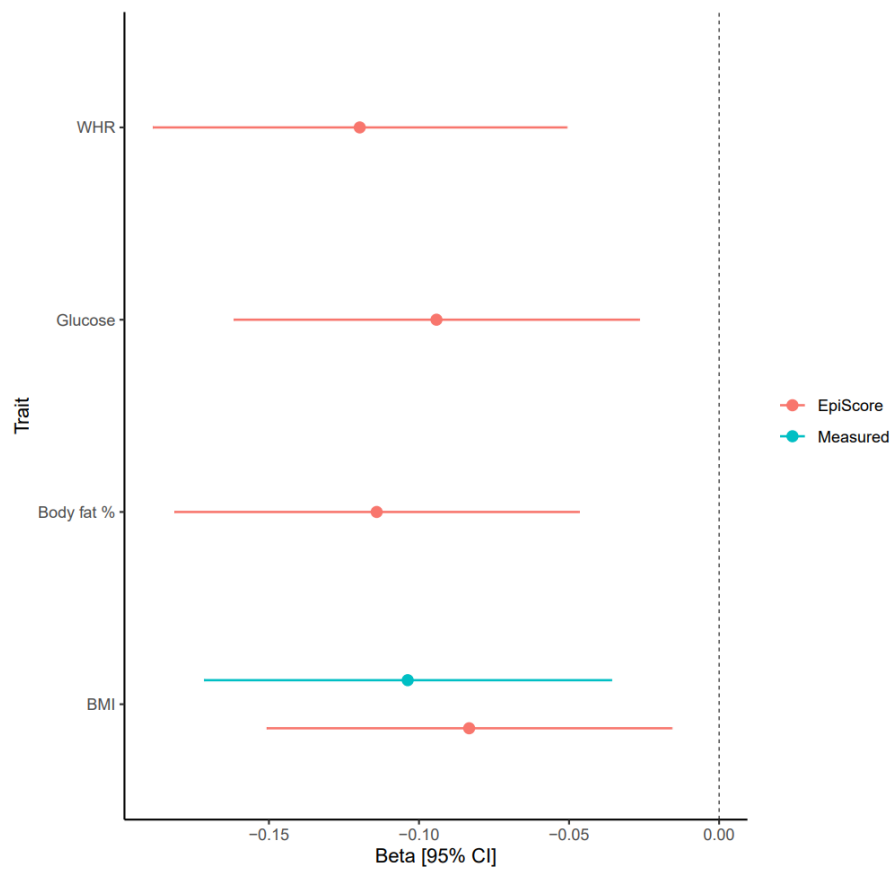
glucose levels were not available for either cohort. Incremental R^2 was calculated for each ancestry group and in the whole cohort in HELIOS. BMI = body mass index, WHR = waist-hip ratio, HDL = high-density lipoprotein.

Next, we tested the Bayesian EpiScores in both LBC1936 and HELIOS, observing similar results to the elnet approach (**Supplementary Figure 8, Supplementary Table 13**).

EpiScore associations with general cognitive function

Metabolic traits have previously been linked to cognitive outcomes. Given this, we tested if the metabolic (elnet) EpiScores were associated with general cognitive function level and longitudinal changes in the LBC1936 (n=861). In models adjusting for age and sex, the three measured traits (BMI, total cholesterol and HDL cholesterol) and all EpiScores, except the total cholesterol EpiScore, were significantly associated with general cognitive function (intercept) in LBC1936 ($P_{FDR} < 0.05$, **Supplementary Figure 9, Supplementary Table 14**). In fully-adjusted models, significant ($P_{FDR} < 0.05$) EpiScore associations were observed for WHR, glucose, body fat percentage and BMI (standardized β_{range} -0.08 to -0.12), and for measured BMI (standardized β : -0.10, **Figure 3A**). No significant associations were observed with general cognitive change over ~12 years (mean age 70 to mean age 82) of follow-up ($P_{FDR} > 0.05$, **Supplementary Table 14**). A combination of EpiScore and measured trait accounted for more variance in general cognitive function level than EpiScore or measured trait alone (**Figure 3B, Supplementary Table 15**). EpiScores augmented the measured trait variance explained for general cognitive function by an average of 0.3%.

A)



B)

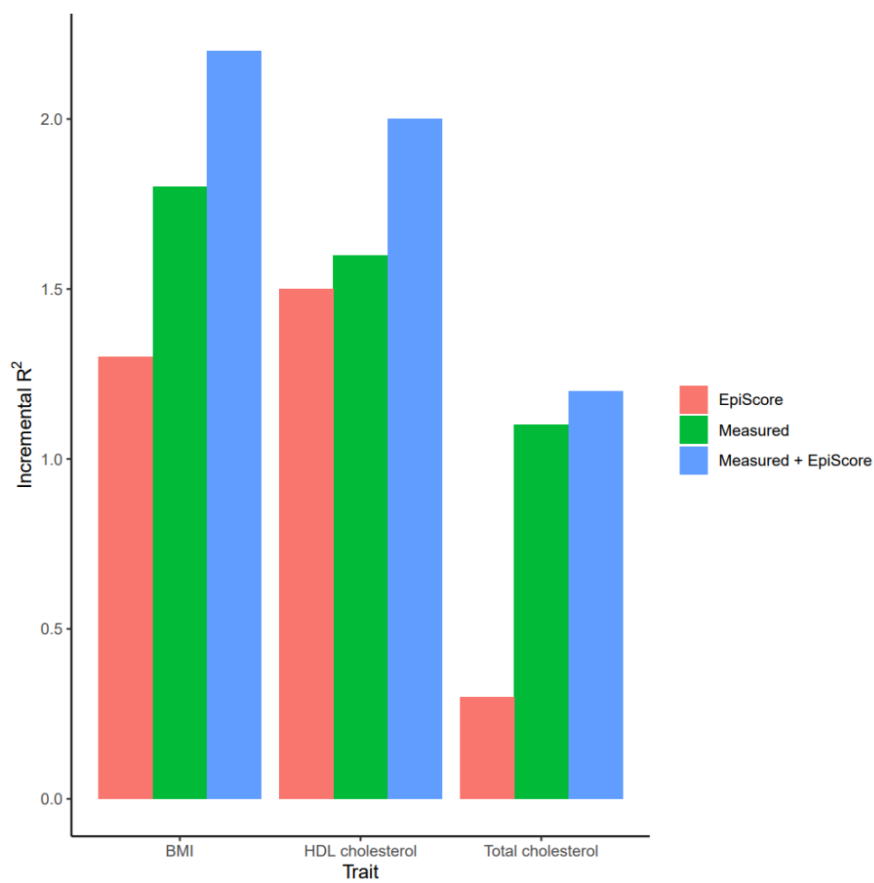


Figure 3: EpiScore and measured metabolic trait in relation to general cognitive function level in the Lothian Birth Cohort 1936 (LBC1936). Panel A shows the significant ($P_{FDR} < 0.05$) associations between measured traits/EpiScores and general cognitive function level in models with full adjustments. Error bars represent 95% confidence intervals. Panel B shows the additional variance accounted for in general cognitive function level by measured metabolic traits, metabolic EpiScores and both combined, over and above linear regression models adjusted for age and sex.

Discussion

Epigenome-wide association studies of six metabolic traits were performed in Generation Scotland ($N > 17,303$). A large number of significantly associated CpGs were identified for each trait via linear regression (marginal associations with $P < 3.6 \times 10^{-8}$ ranged from 304 to 11,815 per trait). A Bayesian approach, which modelled the CpGs jointly and conditionally upon each other, resulted in between 3 and 27 high confidence ($PIP \geq 0.95$) CpG associations for the six traits. EpiScores for each metabolic trait were trained in GS and projected into two independent test cohorts, LBC1936 and HELIOS. The metabolic EpiScores were tested for associations with general cognitive function level and change. Four of the EpiScores were associated with general cognitive function in fully adjusted models ($P_{FDR} < 0.05$), but none were associated with longitudinal cognitive change.

37 CpGs were associated with all six traits when using the marginal linear regression modelling approach. This included 14 CpGs previously linked to metabolic traits in the literature (17-24, 32, 51-54). Gene ontology analysis revealed the genes that the overlapping CpGs mapped to were enriched for relevant biological functions, including regulation of lipid storage, and cholesterol biosynthetic process. Several genes the 37 CpGs mapped to had known metabolic functions. *ABCG1* and *ABCA1* are part of the ABC transporter superfamily involved in the transport of cholesterol (55, 56). *CPT1A*

is a rate-limiting fatty acid oxidation enzyme that oxidises medium and long acyl-CoA esters, an important step that allows these molecules access to the inner mitochondrial membrane (57). *PDK4* is a kinase that inhibits the pyruvate dehydrogenase complex (PDC) which is responsible for the decarboxylation of pyruvate to acetyl-CoA (58). The inhibition of *PDC* results in a switch from glucose oxidation to fatty-acid oxidation and *PDK4* has been suggested as a marker for increased fatty-acid oxidation (58, 59).

Metabolic EpiScores accounted for additional variance in metabolic traits over and above age and sex in both LBC1936 and HELIOS. The elnet EpiScores for BMI and total cholesterol accounted for more variance in their corresponding measured traits in the HELIOS full cohort than in the LBC1936. Conversely, the EpiScore for HDL cholesterol accounted for more variance in the LBC1936 than in the HELIOS full cohort. The performance of elnet metabolic EpiScores in HELIOS varied by ancestry group. In particular, the body fat percentage EpiScore performed similarly in Chinese and Malay individuals (~9% variance accounted for) but had a much lower performance in Indian participants (3.1% variance accounted for). Within the Asian population, it has been reported that Indians have a higher body fat percentage compared with Chinese and Malay populations (60). Asian Indian individuals also have been shown to have increased total and centrally distributed body fat compared with those of European ancestries (61).

The potential usefulness of using DNA methylation to impute measured traits in studies where they are not available was highlighted by the similarity of effect sizes between metabolic EpiScores and their corresponding measured traits in models predicting general cognitive function level (basic adjustments).

This study has multiple strengths including large sample sizes, the use of multi-ancestry cohorts, a multi-method approach (linear regression and Bayesian penalised regression), volunteers from a wide range of ages across adulthood, and longitudinal data to test for cognitive changes in late-life testing (LBC1936). Of the two EWAS strategies, and despite adjustments for relevant covariates, the marginal linear regression approach yielded a vast number of significant CpGs associated with each metabolic trait. However, this approach is naïve in that it does not account for the genome-wide correlation patterns and structure across the methylome. This leads to an inflation in the number of significant findings and biased estimation of effect sizes. Using more stringent methods like BayesR+ helped to overcome such issues, resulting in a high confidence set of CpG-trait associations. Another key strength of the study is that the metabolic EpiScores trained in a cohort of individuals residing in Scotland could account for variance in metabolic traits in a multi-ancestry cohort of Chinese, Malay and Indian Singaporeans. A limitation is that only three of the six metabolic traits were measured in LBC1936, therefore we were unable to compare EpiScore performance against measured WHR, glucose and body fat percentage in this cohort. Finally, alternative strategies for feature pre-selection prior to training EpiScores are likely to result in improved predictors (62, 63).

To conclude, our findings suggest that different EWAS strategies (i.e., marginal linear models and conditional Bayesian models) vastly alter the number of significant CpGs associated with metabolic traits. As increasingly large cohorts with DNA methylation are generated, conditional analyses will help to control false positive rates although they will not identify all correlated/co-dependent sites under a peak. We have also shown that metabolic EpiScores trained in a Scottish population perform well in external Scottish and multi-ancestry Singaporean cohorts. However, further testing is

required in e.g., populations from African or Hispanic ancestries to determine how well the predictors generalise. Further, metabolic EpiScores and measured metabolic traits had comparable magnitudes of association with general cognitive function. This highlights the potential usefulness of metabolic EpiScores to “impute” the corresponding traits where they have not been measured in a cohort.

Availability of data and material

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

Lothian Birth Cohort data are available on request from the Lothian Birth Cohort Study, University of Edinburgh (<https://www.ed.ac.uk/lothian-birth-cohorts/data-access-collaboration>). Lothian Birth Cohort data are not publicly available due to them containing information that could compromise participant consent and confidentiality.

HELIOS data are available on request from the study’s principal investigators. Data access requests for this study should be directed to helios_science@ntu.ed.sg

All code associated with this manuscript is available for open access at the following GitHub repository: https://github.com/hmsmith22/metabolic_trait_project

EWAS summary statistics will be submitted to the EWAS Catalog and Edinburgh DataShare upon publication.

Funding

This research was funded in whole, or in part, by the Wellcome Trust (218493/Z/19/Z, 104036/Z/14/Z, 108890/Z/15/Z, and 221890/Z/20/Z). For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. GS received core support from the Chief Scientist Office of the Scottish Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006). DNA methylation profiling of the GS samples was carried out by the Genetics Core Laboratory at the Edinburgh Clinical Research Facility, Edinburgh, Scotland, and was funded by the Medical Research Council UK and Wellcome (Wellcome Trust Strategic Award STratifying Resilience and Depression Longitudinally (STRADL; Reference 104036/Z/14/Z). DNA methylation data for Generation Scotland was also funded by a 2018 NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation (Ref: 27404; awardee: Dr David M Howard) and by a John, Margaret, Alfred and Stewart Sim Fellowship from the Royal College of Physicians of Edinburgh (Awardee: Dr Heather C Whalley). This work was supported by the European Union Horizon 2020 (PHC.03.15, project No 666881), SVDs@Target, the Fondation Leducq Transatlantic Network of Excellence for the Study of Perivascular Spaces in Small Vessel Disease [ref no. 16 CVD 05]. We thank the LBC1936 participants and team members who contributed to these studies. The LBC1936 is supported by the Biotechnology and Biological Sciences Research Council, and the Economic and Social Research Council [BB/W008793/1] (which supports S.E.H., J.C. and A.T.), Age UK (Disconnected Mind project), the Milton Damerel Trust, the Medical Research Council (G0701120, G1001245, MR/M013111/1, MR/R024065/1) and the University of Edinburgh. Methylation typing of LBC1936 was supported by the 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. H.M.S and D.A.G 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 [218493/Z/19/Z,108890/Z/15/Z]. S.R.C. was supported by a National Institutes of Health (NIH) research grant R01AG054628 and is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 221890/Z/20/Z). D.L.Mc.C. and R.E.M. are supported by Alzheimers Research UK major project grant ARUK/PG2017B/10. E.B and R.E.M. are supported by Alzheimer's Society major project grant AS-PG-19b-010. R.F.H is supported by an MRC IEU Fellowship. The HELIOS study is supported by Singapore Ministry of Health's (MOH) National Medical Research Council (NMRC) under its OF-LCG funding scheme (MOH-000271-00), Singapore Translational Research (StaR) funding scheme (NMRC/StaR/0028/2017), the National Research Foundation, Singapore through the Singapore MOH NMRC and the Precision Health Research, Singapore (PRECISE) under the National Precision Medicine programme (NMRC/PRECISE/2020) and intramural funding from Nanyang Technological University, Lee Kong Chian School of Medicine and the National Healthcare Group.

Ethics approval and consent to participate

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

Ethical approval for the LBC1936 study was obtained from the Multi-Centre Research Ethics Committee for Scotland (Wave 1, MREC/01/0/56) and the Lothian Research Ethics committee (Wave 1, LREC/2003/2/29) and the Scotland A Research Ethics Committee (Waves 2-5, 07/MRE00/58). All participants provided written informed consent. These studies were performed in accordance with the Helsinki declaration.

The HELIOS study was approved by the National Technological University (NTU) Institutional Review Board [IRB-2016-11-030], with written informed consent obtained from each participant before the commencement of the study.

Consent for publication

Not applicable

Competing interests

R.E.M has received a speaker fee from Illumina, is an advisor to the Epigenetic Clock Development Foundation and Optima Partners Ltd. D.A.G and D.L.M. are employed by Optima Partners Ltd in a part-time capacity. R.F.H has acted as a scientific consultant to Optima Partner Ltd and has received consultant fees from Illumina. The remaining authors declare no competing interests.

1. Khan SS, Ning H, Wilkins JT, Allen N, Carnethon M, Berry JD, et al. Association of Body Mass Index With Lifetime Risk of Cardiovascular Disease and Compression of Morbidity. *JAMA Cardiol.* 2018;3(4):280-7.
2. Alloubani A, Nimer R, Samara R. Relationship between Hyperlipidemia, Cardiovascular Disease and Stroke: A Systematic Review. *Curr Cardiol Rev.* 2021;17(6):e051121189015.
3. Salinero-Fort MA, Andrés-Rebollo FJS, Cárdenas-Valladolid J, Méndez-Bailón M, Chico-Moraleja RM, de Santa Pau EC, et al. Cardiovascular risk factors associated with acute myocardial infarction and stroke in the MADIABETES cohort. *Scientific Reports.* 2021;11(1):15245.
4. Cao Q, Yu S, Xiong W, Li Y, Li H, Li J, et al. Waist-hip ratio as a predictor of myocardial infarction risk: A systematic review and meta-analysis. *Medicine (Baltimore).* 2018;97(30):e11639.

5. Wang X, Dong Y, Qi X, Huang C, Hou L. Cholesterol levels and risk of hemorrhagic stroke: a systematic review and meta-analysis. *Stroke*. 2013;44(7):1833-9.
6. Karlsson IK, Gatz M, Arpawong TE, Dahl Aslan AK, Reynolds CA. The dynamic association between body mass index and cognition from midlife through late-life, and the effect of sex and genetic influences. *Scientific Reports*. 2021;11(1):7206.
7. Crane BM, Nichols E, Carlson MC, Deal JA, Gross AL. Body Mass Index and Cognition: Associations Across Mid- to Late Life and Gender Differences. *J Gerontol A Biol Sci Med Sci*. 2023;78(6):988-96.
8. Liu Z, Yang H, Chen S, Cai J, Huang Z. The association between body mass index, waist circumference, waist-hip ratio and cognitive disorder in older adults. *J Public Health (Oxf)*. 2019;41(2):305-12.
9. Liu X, Chen X, Hou L, Xia X, Hu F, Luo S, et al. Associations of Body Mass Index, Visceral Fat Area, Waist Circumference, and Waist-to-Hip Ratio with Cognitive Function in Western China: Results from WCHAT Study. *The journal of nutrition, health & aging*. 2021;25(7):903-8.
10. Crichton GE, Elias MF, Davey A, Sullivan KJ, Robbins MA. Higher HDL cholesterol is associated with better cognitive function: the Maine-Syracuse study. *J Int Neuropsychol Soc*. 2014;20(10):961-70.
11. Svensson T, Sawada N, Mimura M, Nozaki S, Shikimoto R, Tsugane S. The association between midlife serum high-density lipoprotein and mild cognitive impairment and dementia after 19 years of follow-up. *Translational Psychiatry*. 2019;9(1):26.
12. Pang K, Liu C, Tong J, Ouyang W, Hu S, Tang Y. Higher Total Cholesterol Concentration May Be Associated with Better Cognitive Performance among Elderly Females. *Nutrients*. 2022;14(19).
13. Adab P, Pallan M, Whincup PH. Is BMI the best measure of obesity? *BMJ*. 2018;360:k1274.
14. Romero-Corral A, Somers VK, Sierra-Johnson J, Jensen MD, Thomas RJ, Squires RW, et al. Diagnostic performance of body mass index to detect obesity in patients with coronary artery disease. *Eur Heart J*. 2007;28(17):2087-93.
15. Okorodudu DO, Jumean MF, Montori VM, Romero-Corral A, Somers VK, Erwin PJ, et al. Diagnostic performance of body mass index to identify obesity as defined by body adiposity: a systematic review and meta-analysis. *International Journal of Obesity*. 2010;34(5):791-9.
16. Battram T, Yousefi P, Crawford G, Prince C, Sheikhal Babaei M, Sharp G, et al. The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res*. 2022;7:41.
17. Pfeiffer L, Wahl S, Pilling LC, Reischl E, Sandling JK, Kunze S, et al. DNA methylation of lipid-related genes affects blood lipid levels. *Circ Cardiovasc Genet*. 2015;8(2):334-42.
18. Kriebel J, Herder C, Rathmann W, Wahl S, Kunze S, Molnos S, et al. Association between DNA Methylation in Whole Blood and Measures of Glucose Metabolism: KORA F4 Study. *PLoS One*. 2016;11(3):e0152314.
19. Justice AE, Chittoor G, Gondalia R, Melton PE, Lim E, Grove ML, et al. Methylome-wide association study of central adiposity implicates genes involved in immune and endocrine systems. *Epigenomics*. 2020;12(17):1483-99.
20. Sayols-Baixeras S, Subirana I, Fernández-Sanlés A, Sentí M, Lluís-Ganella C, Marrugat J, et al. DNA methylation and obesity traits: An epigenome-wide association study. The REGICOR study. *Epigenetics*. 2017;12(10):909-16.
21. Demerath EW, Guan W, Grove ML, Aslibekyan S, Mendelson M, Zhou YH, et al. Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Hum Mol Genet*. 2015;24(15):4464-79.
22. Braun KVE, Dhana K, de Vries PS, Voortman T, van Meurs JBJ, Uitterlinden AG, et al. Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. *Clin Epigenetics*. 2017;9:15.
23. 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(7635):81-6.

24. Aslibekyan S, Demerath EW, Mendelson M, Zhi D, Guan W, Liang L, et al. Epigenome-wide study identifies novel methylation loci associated with body mass index and waist circumference. *Obesity (Silver Spring)*. 2015;23(7):1493-501.
25. Geurts YM, Dugué PA, Joo JE, Makalic E, Jung CH, Guan W, et al. Novel associations between blood DNA methylation and body mass index in middle-aged and older adults. *International Journal of Obesity*. 2018;42(4):887-96.
26. Sharp GC, Alfano R, Ghantous A, Urquiza J, Rifas-Shiman SL, Page CM, et al. Paternal body mass index and offspring DNA methylation: findings from the PACE consortium. *Int J Epidemiol*. 2021;50(4):1297-315.
27. Vehmeijer FOL, Küpers LK, Sharp GC, Salas LA, Lent S, Jima DD, et al. DNA methylation and body mass index from birth to adolescence: meta-analyses of epigenome-wide association studies. *Genome Med*. 2020;12(1):105.
28. Liu J, Carnero-Montoro E, van Dongen J, Lent S, Nedeljkovic I, Ligthart S, et al. An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis. *Nat Commun*. 2019;10(1):2581.
29. Lim IY, Lin X, Teh AL, Wu Y, Chen L, He M, et al. Dichotomy in the Impact of Elevated Maternal Glucose Levels on Neonatal Epigenome. *J Clin Endocrinol Metab*. 2022;107(3):e1277-e92.
30. Antoun E, Kitaba NT, Titcombe P, Dalrymple KV, Garratt ES, Barton SJ, et al. Maternal dysglycaemia, changes in the infant's epigenome modified with a diet and physical activity intervention in pregnancy: Secondary analysis of a randomised control trial. *PLoS Med*. 2020;17(11):e1003229.
31. Ouidir M, Zeng X, Workalemahu T, Shrestha D, Grantz KL, Mendola P, et al. Early pregnancy dyslipidemia is associated with placental DNA methylation at loci relevant for cardiometabolic diseases. *Epigenomics*. 2020;12(11):921-34.
32. Hedman Å K, 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(1):e001487.
33. McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, et al. Epigenetic prediction of complex traits and death. *Genome Biol*. 2018;19(1):136.
34. Stevenson AJ, McCartney DL, Hillary RF, Campbell A, Morris SW, Bermingham ML, et al. Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clinical Epigenetics*. 2020;12(1):113.
35. Green C, Shen X, Stevenson AJ, Conole ELS, Harris MA, Barbu MC, et al. Structural brain correlates of serum and epigenetic markers of inflammation in major depressive disorder. *Brain Behav Immun*. 2021;92:39-48.
36. Hamilton OKL, Zhang Q, McRae AF, Walker RM, Morris SW, Redmond P, et al. An epigenetic score for BMI based on DNA methylation correlates with poor physical health and major disease in the Lothian Birth Cohort. *Int J Obes (Lond)*. 2019;43(9):1795-802.
37. Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM, et al. Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *International Journal of Epidemiology*. 2013;42(3):689-700.
38. Taylor AM, Pattie A, Deary IJ. Cohort Profile Update: The Lothian Birth Cohorts of 1921 and 1936. *International Journal of Epidemiology*. 2018;47(4):1042-r.
39. Deary IJ, Gow AJ, Pattie A, Starr JM. Cohort profile: the Lothian Birth Cohorts of 1921 and 1936. *Int J Epidemiol*. 2012;41(6):1576-84.
40. Zhang F, Chen W, Zhu Z, Zhang Q, Nabais MF, Qi T, et al. OSCA: a tool for omic-data-based complex trait analysis. *Genome Biology*. 2019;20(1):107.
41. Therneau JSaT. kinship2: Pedigree Functions. 2022.
42. Therneau TM. coxme: Mixed Effects Cox Models. 2022.

43. Bollepalli S, Korhonen T, Kaprio J, Anders S, Ollikainen M. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. *Epigenomics*. 2019;11(13):1469-86.
44. Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. *PLOS ONE*. 2013;8(5):e63812.
45. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;13:86.
46. Saffari A, Silver MJ, Zavattari P, Moi L, Columbano A, Meaburn EL, et al. Estimation of a significance threshold for epigenome-wide association studies. *Genet Epidemiol*. 2018;42(1):20-33.
47. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *ArXiv*. 2011;abs/1201.0490.
48. Maksimovic J, Oshlack A, Phipson B. Gene set enrichment analysis for genome-wide DNA methylation data. *Genome Biology*. 2021;22(1):173.
49. Trejo Banos D, McCartney DL, Patxot M, Anchieri L, Battram T, Christiansen C, et al. Bayesian reassessment of the epigenetic architecture of complex traits. *Nature Communications*. 2020;11(1):2865.
50. Rosseel Y. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*. 2012;48(2):1 - 36.
51. Mendelson MM, Marioni RE, Joehanes R, Liu C, Hedman Å K, Aslibekyan S, et al. Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach. *PLoS Med*. 2017;14(1):e1002215.
52. Shah S, Bonder MJ, Marioni RE, Zhu Z, McRae AF, Zhernakova A, et al. Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations. *Am J Hum Genet*. 2015;97(1):75-85.
53. Sun D, Zhang T, Su S, Hao G, Chen T, Li QZ, et al. Body Mass Index Drives Changes in DNA Methylation: A Longitudinal Study. *Circ Res*. 2019;125(9):824-33.
54. Geurts YM, Dugué PA, Joo JE, Makalic E, Jung CH, Guan W, et al. Novel associations between blood DNA methylation and body mass index in middle-aged and older adults. *Int J Obes (Lond)*. 2018;42(4):887-96.
55. Matsuo M. ABCA1 and ABCG1 as potential therapeutic targets for the prevention of atherosclerosis. *Journal of Pharmacological Sciences*. 2022;148(2):197-203.
56. Kobayashi A, Takanezawa Y, Hirata T, Shimizu Y, Misasa K, Kioka N, et al. Efflux of sphingomyelin, cholesterol, and phosphatidylcholine by ABCG1. *Journal of Lipid Research*. 2006;47(8):1791-802.
57. Schlaepfer IR, Joshi M. CPT1A-mediated Fat Oxidation, Mechanisms, and Therapeutic Potential. *Endocrinology*. 2020;161(2).
58. Sugden MC, Holness MJ. Mechanisms underlying regulation of the expression and activities of the mammalian pyruvate dehydrogenase kinases. *Archives of Physiology and Biochemistry*. 2006;112(3):139-49.
59. Pettersen IKN, Tusubira D, Ashrafi H, Dyrstad SE, Hansen L, Liu X-Z, et al. Upregulated PDK4 expression is a sensitive marker of increased fatty acid oxidation. *Mitochondrion*. 2019;49:97-110.
60. Wulan SN, Westerterp KR, Plasqui G. Ethnic differences in body composition and the associated metabolic profile: A comparative study between Asians and Caucasians. *Maturitas*. 2010;65(4):315-9.
61. Freitas I, Plank LD, Rush EC. Body size, body composition and fat distribution: comparative analysis of European, Maori, Pacific Island and Asian Indian adults. *British Journal of Nutrition*. 2009;102(4):632-41.
62. Merzbacher C, Ryan B, Goldsborough T, Hillary RF, Campbell A, Murphy L, et al. Integration of datasets for individual prediction of DNA methylation-based biomarkers. *Genome Biology*. 2023;24(1):278.
63. Cheng Y, Gieger C, Campbell A, McIntosh A, Waldenberger M, McCartney D, et al. Feature pre-selection for the development of epigenetic biomarkers. *medRxiv*; 2024.

