Elliott et al. Clinical Epigenetics (2022) 14:130
https://doi.org/10.1186/s13148-022-01351-2

Clinical Epigenetics

RESEARCH

Open Access

Characterisation of ethnic differences in DNA methylation between UK-resident South Asians and Europeans

Hannah R. Elliott[1,2*], Kimberley Burrows[1,2], Josine L. Min[1,2], Therese Tillin[3,4], Dan Mason[5], John Wright[5], Gillian Santorelli[5], George Davey Smith[1,2], Deborah A. Lawlor[1,2], Alun D. Hughes[3,4], Nishi Chaturvedi[3,4] and Caroline L. Relton[1,2]

Abstract

Ethnic differences in non-communicable disease risk have been described between individuals of South Asian and European ethnicity that are only partially explained by genetics and other known risk factors. DNA methylation is one underexplored mechanism that may explain differences in disease risk. Currently, there is little knowledge of how DNA methylation varies between South Asian and European ethnicities. This study characterised differences in blood DNA methylation between individuals of self-reported European and South Asian ethnicity from two UK-based cohorts: Southall and Brent Revisited and Born in Bradford. DNA methylation differences between ethnicities were widespread throughout the genome ($n = 16{,}433$ CpG sites, 3.4% sites tested). Specifically, 76% of associations were attributable to ethnic differences in cell composition with fewer effects attributable to smoking and genetic variation. Ethnicity-associated CpG sites were enriched for EWAS Catalog phenotypes including metabolites. This work highlights the need to consider ethnic diversity in epigenetic research.

Keywords: DNA methylation, Epigenetics, mQTLs, Ethnicity, Ancestry

Introduction

There are well-described ethnic differences in non-communicable diseases between migrants of South Asian genetic ancestry and their European ancestry counterparts [1, 2]. The most striking difference is the greater risk of cardiometabolic disease in South Asians compared to Europeans (reviewed [3, 4]). In contrast, all-cancer morbidity and mortality risks are lower in South Asian migrant groups compared to Europeans [2, 5]. Ethnic differences in non-communicable diseases are only partially influenced by known genetic or environmental factors [4, 6–9]. Although beyond the scope of this paper, it is possible that the colonial history of UK South Asian people and persistent racism could affect their increased prevalence of chronic diseases such as type 2 diabetes and coronary heart disease, for example through psychosocial paths [10, 11] and through inequitable treatment within health services [12]. Focussing on molecular mechanisms, DNA methylation is one of the underexplored epigenetic mechanisms that may explain differences in disease risk.

In recent years, Epigenome-Wide Association Studies (EWASs) have associated DNA methylation with diverse health-related outcomes such as cancer [13–16], diabetes [17–19], rheumatoid arthritis [20], psychosis and schizophrenia [21], blood pressure [22, 23], circulating metabolic measures [24] and BMI [25–28]. DNA methylation has also proved a useful biomarker of lifestyle exposures, for example smoking [29–32] and

*Correspondence: hannah.elliott@bristol.ac.uk

[1] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK
Full list of author information is available at the end of the article

alcohol intake [33, 34]. Furthermore, DNA methylation may also be a predictor of future disease, for example, myocardial infarction and coronary heart disease [35] or all-cause mortality and time to death [36, 37]. Ethnic differences have been explored in the context of these EWAS. For example, a meta-EWAS of 1250 incident type 2 diabetes cases and 1950 controls drawn from five European cohorts identified 76 CpG sites associated with incident type 2 diabetes [38]. Of these findings, 64 (84.2%) were directionally consistent in an independent cohort of Indian Asians. The evidence therefore suggests that many but not all DNA methylation associations with exposures or outcomes are consistent across populations indicating a possible role for DNA methylation as a mediator of excess disease risk between population groups.

Studies have begun to identify ethnic differences in DNA methylation patterns across the genome [39–41]. These studies highlight that ethnic differences in DNA methylation are likely to be frequent in the genome and Galanter et al. [39] suggested they arise as a consequence of both genetic and environmental factors. However, to our knowledge there has been no large-scale genome-wide comparison of DNA methylation in South Asian and European ethnic groups.

The aim of this study was therefore to identify and characterise DNA methylation differences between healthy adult individuals of self-reported South Asian and European ethnicity resident in the UK. We did this by first identifying cross-sectional differences in genome-wide DNA methylation and then assessing the stability of signals over time. We then measured ethnic differences in DNA methylation-derived estimates of cell composition and the methylation-derived neutrophil-to-lymphocyte ratio (mdNLR), an index of systemic inflammation. We report the relative contribution of genetic and environmental sources of variance in ethnicity-associated DNA methylation and also explore sources of environmental variation. We finally assess whether ethnicity-associated CpG sites are functionally implicated in biological pathways (using publicly available databases) or disease endpoints or risk factors captured by the EWAS Catalog [42].

Analyses were conducted in the Southall And Brent Revisited (SABRE) cohort [3] and replicated in the Born in Bradford (BiB) cohort [43, 44]. These cohorts both include individuals of south Asian and Europeans in the UK, but they are distinct in that they recruited from different geographical areas of the UK with South Asians in SABRE predominantly of Indian origin and South Asians in BiB predominantly of Pakistani origin. SABRE and BiB also differ in age, sex and other exposures, for example, smoking behaviours. The inclusion of these two distinct cohorts is a strength of the study design, allowing the identification of ethnicity-associated DNA methylation sites that are likely to be generalisable.

Results

Cohort characteristics

A comparison of cohort and subgroup characteristics is shown in Table 1. All SABRE participants were male and recruited in middle age [3]. BiB participants were female, with DNA methylation measured on samples collected in pregnancy (24–28 weeks of gestation) [43, 44]. Distributions of age, BMI and smoking differed between the two studies and for some measures, between the two ethnic groups (see Table 1).

Concordance of self-reported ethnicity with genetic ancestry

Principal components analysis (PCA) of SABRE, BiB and HapMap3 populations was used to assign individuals to groups with similar genetic ancestry (Additional file 1: Fig. S1).

Self-reported Europeans from SABRE and BiB cluster closely with Europeans from HapMap3 (Additional file 1: Fig. S1). Correspondingly, self-reported South Asians from SABRE and BiB cluster with or near HapMap3 Gujarati Indians recruited from Houston, USA. PCA analysis on the self-reported South Asian and European individuals in SABRE and BiB show separate clustering by ethnicity. The genetic variation as estimated by genetic PCs appears higher in the South Asian group, where ethnicity subgroups can be separated on both the PC1 and PC2 axis. In SABRE, those indicating their ethnicity subgrouping as "Punjabi" or "Gujarati" cluster more closely together than other subgroups.

Concordance of self-reported ethnicity with DNA methylation components

Principal components (PCs) were derived from the SABRE and BiB methylation matrices. When utilising the 1000 or 10,000 probes with the greatest variance, PC1 differentiated between South Asian and European individuals in both SABRE and BiB (Additional file 2: Fig. S2). There was no differential clustering by the ethnic group when using all independent probes from the array ($n_{SABRE} = 21{,}023$, $n_{BiB} = 19{,}438$).

Identification of ethnic differences in DNA methylation

Ethnicity EWAS

In univariable analysis, differential methylation between SABRE individuals of self-reported South Asian or European ethnicity was identified at 32,435 CpG sites at $p \le 1.03 \times 10^{-7}$ (6.7% of the 484,781 sites assessed, Additional file 3: Table S1). Effect sizes ranged from 0.08 to 25.8% difference in DNA methylation between ethnic

Table 1 Cohort characteristics

Variable	SABRE baseline			SABRE follow-up			Born in Bradford			Cross-cohort		
	European mean (SD) or n (%)	South Asian mean (SD) or n (%),	p value[a]	European mean (SD) or n (%)	South Asian mean (SD) or n (%),	p value[a]	White British mean (SD) or n (%)	Asian Pakistani mean (SD) or n (%),	p value[a]	SABRE baseline	BiB	p value[b]
N (%)	400 (50.0)	400 (50.0)		66 (47.5)	73 (52.5)		444 (48.5)	472 (51.5)		800 (46.6)	916 (53.4)	
Males (%)	400 (100)	400 (100)		66 (100)	73 (100)		0 (0)	0 (0)		800 (100)	0 (0)	
Age (years)	52.3 (7.1)	51.0 (7.1)	0.084	68.6 (5.3)	69.0 (5.8)	0.634	26.8 (6.1)	28 (5.3)	0.00205	51.9 (7.2)	27 (5.7)	0
BMI (kg/m^2)	26.1 (3.7)	26.0 (3.6)	0.441	26.7 (3.1)	26 (4.0)	0.318	27.1 (6.3)	26 (5.2)	0.00269	26 (3.6)	26 (5.8)	0.0542
Bcell (%)	7.6 (2.09)	9.1 (2.69)	9.58E−20	8.3 (2.78)	10.0(6.84)	0.0424	4.75 (1.2)	5.7 (1.5)	1.62E−27	8.35 (2.5)	5.3 (1.4)	2.03E−151
CD4T (%)	14.5 (4.84)	14.0 (4.32)	0.00549	16.9 (5.74)	15.0 (6.38)	0.102	11.8 (2.6)	12 (3.0)	0.00104	14 (4.6)	12 (2.8)	1.50E−23
CD8T (%)	1.2 (2.23)	2.1 (3.04)	1.53E−06	0.5 (1.5)	1.6 (2.74)	0.00651	0.126 (0.7)	0.36 (1.1)	0.000122	1.66 (2.7)	0.25 (0.9)	1.24E−40
Eos (%)	0.06(0.368)	0.26 (1.15)	0.000598	0.2 (0.722)	0.7 (1.78)	0.0209	0.0185 (0.2)	0.19 (0.8)	1.57E−05	0.159 (0.9)	0.11 (0.6)	0.154
Mono (%)	10.3 (1.94)	9.9 (1.69)	0.00166	10.7 (2.04)	9.5 (1.67)	0.000175	11.2 (1.6)	12 (2.01)	0.00113	10.1 (1.8)	11 (1.8)	8.13E−45
Neu (%)	56.2 (8.52)	53 (8.47)	3.35E−06	53.9 (10.6)	52 (13.2)	0.308	68.7 (5.0)	65 (6.97)	1.77E−20	54.8 (8.6)	67 (6.4)	1.74E−171
NK (%)	15.8 (4.41)	18 (4.65)	8.97E−09	16 (4.7)	18 (4.8)	0.0227	6.55 (2.4)	8.4 (2.74)	8.51E−26	16.7 (4.6)	7.5 (2.7)	1.78E−295
Ever smoking (%)	291 (72.8)	127 (31.8)	1.60E−30	45 (68.2)	22 (31.1)	1.60E−05	260 (58.6)	35 (7.4)	6.40E−61	418 (52.3)	295 (32.2)	1.80E−17

[a] *t* test or χ^2 for ethnic differences

[b] *t* test or χ^2 comparing SABRE baseline and Born in Bradford

groups. CpG sites associated with ethnicity were distributed throughout the genome (Fig. 1A, B), with hypermethylation amongst individuals of South Asian ethnicity being predominant (67% of CpG sites identified) when compared to Europeans (Fig. 1A, C).

In BiB, effect estimates correlated with those reported in SABRE ($R^2 = 0.88$, see Fig. 2A). When considering p values, 13,803/30,095 (46%) of the individual CpG sites were replicated as defined by the maintenance of p values below the epigenome-wide threshold (Additional file 3: Table S1). We observed inflated λ values in both SABRE ($\lambda = 2.93$) and BiB ($\lambda = 2.22$) EWAS. λ values were recalculated as described previously [45] (see methods). λ was substantially reduced in both SABRE ($\lambda = 1.53$) and BiB ($\lambda = 1.02$), indicating inflation is likely to stem from the abundance of the true biological signal rather than substantial residual confounding.

Differentially methylated regions (DMRs)

DMR analysis was applied to the EWAS results to identify wider regions of DNA methylation that were associated with self-reported ethnicity. We identified 10,675 DMRs containing 2 or more CpGs that differed between the two ethnic groups at adjusted $p < 0.05$ in SABRE (Additional file 4: Table S2). Using the same analysis parameters, 5371 (50%) of DMRs identified in SABRE had at least 1 overlapping CpG with DMRs identified in BiB (Additional file 3: Table S1 and Additional file 4: Table S2). The modal number of CpGs amongst replicated DMRs was 2 ($n = 1870$), and 73% of replicated DMRs contained fewer than 5 CpGs. The largest region identified in the DMR analysis was 1.1 kb in length and contained 31 CpG sites. This region on chromosome 11 spans the upstream and exon 1 region of the imprinted *KCNQ1DN* locus and appears to contain CpG islands, DNase I hypersensitivity clusters, and transcription factor binding sites [46, 47].

All further analyses were restricted to a pool of 16,344 unique CpG sites: 13,803 replicated CpG sites identified in the univariable EWAS of ethnicity in SABRE and BiB and 2541 CpGs representing replicated DMRs (the CpG with the lowest EWAS p value selected from each DMR, Additional file 3: Table S1).

Ethnic differences in DNA methylation are stable over time

Participants in the SABRE study attended a follow-up clinic approximately 20 years after baseline recruitment ($n = 139$), allowing an assessment of the temporal stability of ethnicity-associated DNA methylation variation. Characteristics of individuals from the SABRE follow-up timepoint are shown in Table 1. Ethnicity EWAS effect estimates were compared between SABRE timepoints at replicated CpG sites ($n = 16,344$ CpGs). Effect estimates correlated with those reported at the SABRE baseline

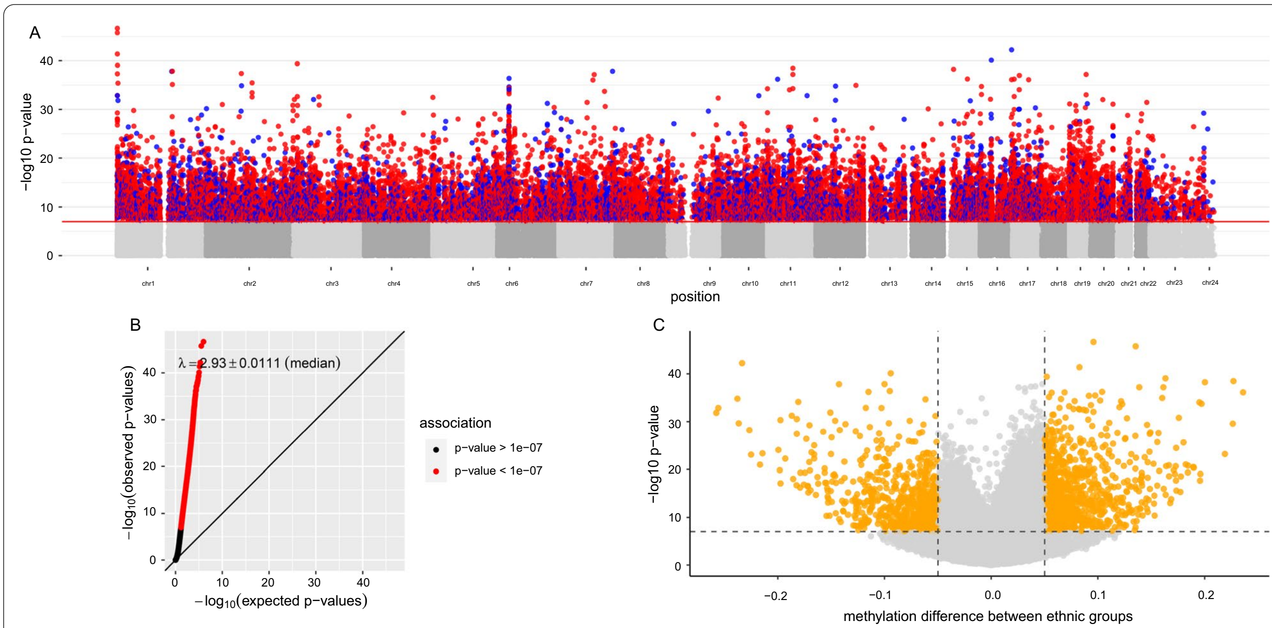

Fig. 1 EWAS plots showing association between methylation and ethnic group in the SABRE cohort. **A** Manhattan plot showing association between ethnicity and DNA methylation. CpG sites with corresponding p values at $\leq 1.03 \times 10^{-7}$ are colour-coded to show the direction of effect: red CpG sites are hypermethylated in South Asian individuals, while blue CpG sites are hypomethylated relative to the European group. **B** *Q–Q* plot showing observed vs expected p values from the EWAS analysis. The red line denotes equality. $\lambda = 2.93$. **C** Volcano plot showing p value versus effect size for each of the tested CpGs on the array. Orange highlighted CpGs are those with $p < 1.03 \times 10^{-7}$ and with effect sizes of $> 5\%$

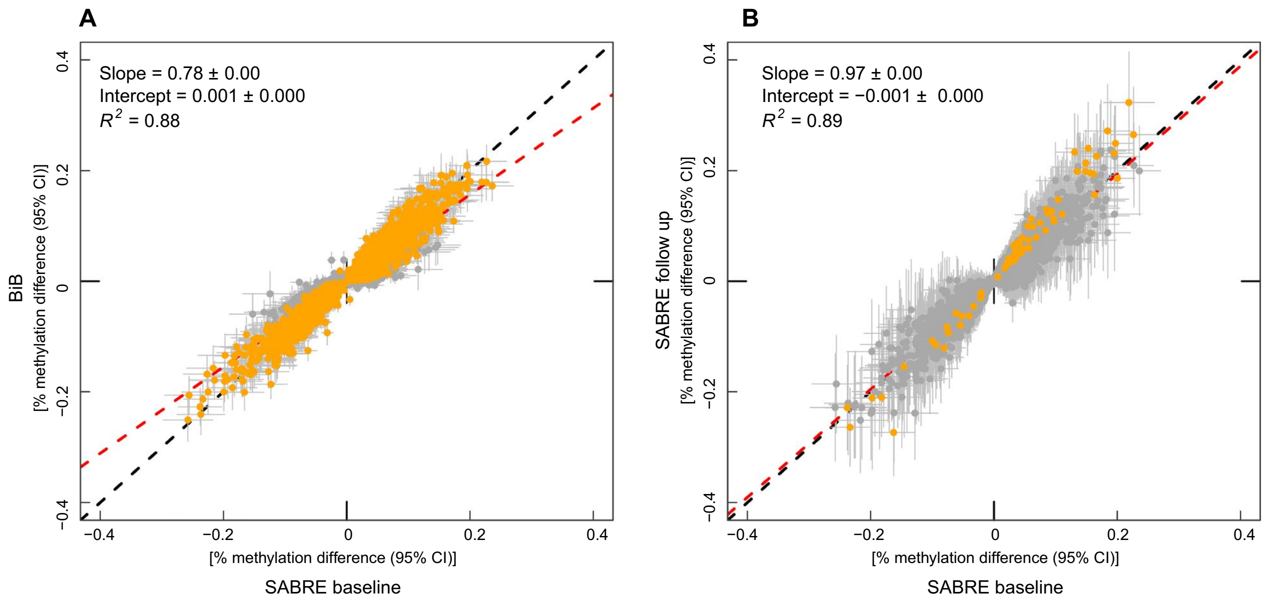

Fig. 2 Correlation of ethnicity EWAS effect estimates between **A** SABRE baseline and BiB individuals **B** SABRE baseline and follow-up timepoints. Each CpG is represented by a point on the graph with 95% confidence intervals for effect estimates. Red dashed line: linear regression between data sets. Black dashed line: line of equality. Orange highlighted estimates: $p \leq 1.03 \times 10^{-7}$ in BiB (**A**) or SABRE follow-up timepoint (**B**) EWAS. **A** $N = 30{,}081$ CpG sites (those associated with ethnicity from SABRE EWAS with data available in BiB. **B** $N = 15{,}131$ CpG sites associated with self-reported ethnicity at SABRE baseline and replicated in BiB

($R^2 = 0.89$, Fig. 2B). Only 1685 (10.3%) CpG sites had effect estimates that changed by > twofold between time points.

Analysis of ethnic differences in methylation-derived phenotypes

DNA methylation data were used to generate an index of biological ageing (using a variety of epigenetic clocks [48]), estimates of white blood cell counts [49] and an index of systemic inflammation [50]. Subsequent analyses assessed whether these derived phenotypes differed between ethnic groups.

Ethnic differences in epigenetic age acceleration

Epigenetic age was computed using five clock algorithms (see Methods). SABRE South Asians had higher age acceleration estimates amongst all clock measures (except the Hannum clock) compared to SABRE Europeans in unadjusted tests (Additional file 4: Table S3). For both Extrinsic and Intrinsic Epigenetic Age Accelerations (EEAA and IEAA), these differences were robust to adjustment for BMI and self-reported smoking in PhenoAge (EEAA: beta = 1.63, $p = 7.62 \times 10^{-5}$; IEAA: beta = 1.59, $p = 6.39 \times 10^{-5}$) and SkinBlood estimators (EEAA: beta = 0.75, $p = 1.50 \times 10^{-3}$; IEAA: beta = 0.75, $p = 9.78 \times 10^{-4}$). Evidence for an association between age acceleration and ethnicity in BiB was weak (Additional file 4: Table S3).

Ethnic differences in cell composition

Using a linear model where estimated cell proportion was the outcome, ethnicity was the predictor and age was included as a covariate (Additional file 4: Table S4, model 1) we show that SABRE and BiB self-reported South Asian individuals had elevated estimated proportions of B cells, CD8 + T cells, eosinophils and natural killer cells and reduced estimated proportions of Neutrophils compared to self-reported European individuals from the same cohort (Table 1). These relationships were unchanged when including smoking status as an additional covariate (Additional file 4: Table S4, model 2). Differences in the proportion of cell subgroups therefore exist between ethnic groups but these differences are not substantially explained by age or smoking status.

Ethnic differences in systemic inflammation

Methylation-derived neutrophil-to-lymphocyte ratio (mdNLR), an index of systemic inflammation, was 13.6% lower in SABRE South Asians compared to Europeans (median (IQR): South Asians = 1.25 (0.70), Europeans = 1.42 (0.73); $p = 2.47 \times 10^{-7}$). This relationship was unchanged by adjustment for age, smoking and BMI. Similar findings were observed in BiB, with mdNLR

18.1% lower in South Asians compared to Europeans (median (IQR): South Asians = 2.48 (1.04), Europeans = 3.03 (1.14); $p = 1.59 \times 10^{-20}$) in univariable analyses, and unchanged with adjustment for age, smoking or BMI.

Determinants of ethnic differences in DNA methylation

To explore the potential factors mediating the relationship between self-reported ethnicity and DNA methylation, we undertook a series of further analyses in SABRE.

The contribution of ethnic differences in smoking, age and cell composition to observed ethnic differences in DNA methylation

There were marked differences in self-reported smoking behaviour between SABRE ethnic groups and European participants were slightly younger than South Asian participants (see Table 1). Age and smoking status have both been shown to be associated with DNA methylation in previous studies [51, 52]. Adjusting the univariable ethnicity EWAS model for age did not alter effect estimates ($R^2 = 1.0$, Additional file 3: Table S1). Effect estimates when comparing the univariable ethnicity EWAS model with a model adjusted for age and self-reported smoking status were also highly correlated ($R^2 = 0.98$, Additional file 3: Table S1). However, the *p* values of 2995/16,344 (18%) of CpG sites were attenuated ($p > 1.03 \times 10^{-7}$) following adjustment for age and smoking status; 9/16,344 (0.05%) of CpG sites had effect estimates that changed by > twofold between models (Additional file 3: Table S1). Ethnicity-associated CpGs were compared to results from a large meta-analysis of smoking in adults [51]. We observed an increased proportion of smoking-associated CpGs amongst the ethnicity-associated CpG sites in SABRE compared to expected (chi-square test: 23.5% vs 7.4%, $p < 2.2e^{-16}$). We therefore conclude that smoking differences are likely to contribute to the observed ethnic differences in DNA methylation in up to 18% of CpG sites tested. Smoking-associated CpGs [51] are annotated in Additional file 3: Table S1.

Cell composition estimates were included as additional covariates in the univariable ethnicity EWAS model. In this analysis, effect estimates were correlated ($R^2 = 0.9$) (Additional file 5: Fig. S3). The *p* values of 12,422/16,344 (76%) of CpG sites were attenuated following adjustment for cell subtypes ($p > 1.03 \times 10^{-7}$) and 4396/16,344 (26.9%) of CpG sites had effect estimates that changed by > twofold between models (Additional file 3: Table S1 Additional file 5: Fig. S3). Additional adjustment for age and smoking status had a minimal effect: *p* values of 12,580/16,344 (77%) of CpG sites were attenuated following adjustment for age, smoking status and cell subtypes combined ($p > 1.03 \times 10^{-7}$) and 3198/16,344 (19.6%) of CpG sites had effect estimates that changed by > twofold between models (Additional file 3: Table S1). Differences in cell composition therefore appear to be an important driver of ethnicity-associated methylation patterns and these differences are not strongly driven by differences in age or smoking status between ethnic groups.

Genetic contribution to ethnic differences in DNA methylation

Assessment of genetic principal components Genetic principal components generated by PCA ($n = 20$, see methods) were included as additional covariates in the univariable ethnicity EWAS model. None of the 16,344 ethnicity-associated CpG sites had *p* values at $p \leq 1.03 \times 10^{-7}$ following adjustment for genetic principal components.

Assessment of mQTLs Using the Genetics of DNA Methylation Consortium (GoDMC) data [53], 16,720 mQTLs were identified (containing 15,566 unique SNPs and 10,171 unique CpGs) [53]. 62% of ethnicity-associated CpGs had at least one mQTL (10,171 of 16,344 CpGs queried, CpGs annotated in Additional file 3: Table S1). In GoDMC, 45% of tested CpGs had an mQTL, ethnicity-associated CpGs in this study therefore appear to be enriched for mQTLs (OR = 1.00, 95%CI = 1.93–2.06, $p = 0$). A total of 8686 SNPs were available in the SABRE cohort which allowed us to further explore 9382 mQTLs across 8318 unique CpGs.

We tested each available SNP to identify differences in allele frequency between SABRE ethnic groups. We estimate that 2908 (33.5%) of the SNPs tested differ in allele frequency between populations (adjusted $p < 0.05$, $n_{tests} = 8686$) (Additional file 4: Table S5, annotated in Additional file 3: Table S1). To investigate further, we included the respective mQTL genotype as a covariate in the univariable ethnicity EWAS model for each CpG with an mQTL (Additional file 4: Table S5). Variance explained (measured by R^2) by the models increased by a mean of 2.7% (SD = 5.3%) in CpGs with differences in allele frequency between populations and a mean of 1.7% (SD = 3.2%) in CpGs showing no differences in allele frequency between populations.

We therefore conclude that ethnic differences in DNA methylation may be partially attributable to differences in allele frequency between ethnicities but overall the contribution of these genetic effects are small. Amongst CpGs with mQTLs that varied in allele frequency between ethnic groups, adjusting for the mQTL attenuated the EWAS signal above the EWAS *p* value threshold in only 763/2908 (26.2%) of unique CpG sites (Additional file 4: Table S5).

We also postulated that DNA methylation differences between ethnic groups may be driven by differences in the mQTL effect between ethnicities. We therefore tested each available SNP to identify interactions between mQTL and ethnic group on methylation using results from the GEM interaction model [54]. GxE models were run where methylation was the outcome variable, genotype x ethnicity was the predictor and age and smoking were included as covariates. We identified 67 interactions (adjusted $p < 0.05$, $n_{\text{tests}} = 9382$) between ethnicity and mQTLs indicating that a small number of CpG associations with ethnicity may be attributable to ethnic-specific genetic effects (Additional file 3: Table S1, Additional file 4: Table S6). None of the 61 unique interaction mQTLs overlaps with a multi-ethnic GWAS of cell composition [55]. We also compared the 66 unique CpG sites showing evidence of interaction between ethnicity and mQTLs with the univariable EWAS and the EWAS model which included cell composition as a covariate. The p values of 15/66 (23%) of CpG sites were attenuated ($p > 1.03 \times 10^{-7}$) following adjustment for cell composition but none of the effect estimates changed by > twofold between models (Additional file 3: Table S1). This indicates that interactions between ethnicity and mQTLs are not related to differences in cell composition.

Functional annotation of ethnicity-associated CpG sites

We initially identified a pool of 16,344 unique CpG sites that were associated with an ethnic group in SABRE and replicated these in BiB. To characterise them in terms of biology and disease relevance we undertook a series of further analyses.

Ethnicity-associated CpGs are depleted for CpGs in 3′UTRs and intronic regions

Ethnicity-associated CpGs were depleted for CpGs in 3′UTRs (OR = 0.71, 95% CI = 0.64–0.78, $p = 1.19 \times 10^{-12}$) and depleted for CpGs in intronic regions (OR = 0.90, 95% CI = 0.87–0.93, $p = 2.32 \times 10^{-10}$) based on ANNOVAR annotation compared to all CpGs on the 450 k array (Additional file 4: Table S7).

Ethnicity-associated CpGs are enriched in multiple phenotype groups based on EWAS Catalog data

Results from the EWAS Catalog [42] were grouped into related phenotypes [56] and tested for enrichment. The 16,344 ethnicity-associated CpGs are strongly enriched for CpG sites previously reported to be associated with metabolites (OR = 4.12, 95%CI = 3.64–4.67, $p = 3.14 \times 10^{-129}$), smoking (OR = 2.20, 95%CI = 2.06–2.35, $p = 1.13 \times 10^{-130}$), alcohol (OR = 2.10, 95%CI = 1.80–2.43, $p = 1.47 \times 10^{-23}$), cancer (OR = 1.83, 95%CI = 1.77–1.90, $p = 4.11 \times 10^{-251}$) and perinatal phenotypes (OR = 1.89, 95%CI = 1.58–2.26, $p = 1.00 \times 10^{-12}$) and strongly depleted for autoimmune (OR = 0.34, 95%CI = 0.32–0.36, $p = 1.31 \times 10^{-278}$) and infection phenotypes (OR = 0.52, 95%CI = 0.48–0.57, $p = 1.02 \times 10^{-52}$) compared to all entries in the EWAS Catalog. CpG sites are also enriched for CpG sites associated with ancestry (OR = 1.50, 95%CI = 1.24–1.82, $p = 1.80 \times 10^{-5}$) (Fig. 3).

Ethnicity-associated CpG sites are highly enriched for cis-eQTMs

In total, 1372 ethnicity-associated CpGs were expression quantitative trait methylation (eQTMs) based on BIOS data [57] and ethnicity-associated CpGs were highly enriched for eQTMs compared to the overall BIOS data set (OR = 2.72, 95% CI = 2.57, 2.88, $p = 6.78 \times 10^{-271}$). eQTMs are annotated in Additional file 3: Table S1.

Ethnicity-associated CpG sites are enriched for several GO terms

We separated ethnicity-associated CpG sites into those with higher or lower levels of methylation in the South Asian compared to the European group. These are referred to as the "higher methylation set" and "lower methylation set". There were 427 GO terms enriched at FDR < 0.05 amongst higher methylation set CpGs. There was enrichment for Biological Process terms including several related to morphogenesis and development. Others were related to synaptic signalling. There was also enrichment for cellular component terms "integral/intrinsic component of plasma membrane", "plasma membrane" and "cell periphery" (Additional file 4: Table S8). There were 50 GO terms enriched at FDR < 0.05 amongst the lower methylation set CpGs. Enrichment of Biological Process terms were predominantly terms related to immune response. Enrichment of Cellular Component terms were related to "cell surface" and "plasma membrane" (Additional file 4: Table S8).

We repeated the GO ontology enrichment analysis, this time restricting to CpGs that were identified as eQTMs based on BIOS data (Additional file 4: Table S8). In this analysis, CpGs in the higher methylation set were enriched for 18 GO terms, almost all were Molecular Function terms. The most strongly enriched terms were "transcription regulatory region sequence-specific DNA binding", "RNA polymerase II transcription regulatory region sequence-specific DNA binding", "RNA polymerase II cis-regulatory region sequence-specific DNA binding" and "regulatory region nucleic acid binding". Amongst CpGs in the lower methylation set, we detected enrichment of 54 GO terms predominantly of Biological Process terms relating to immune response (Additional file 4: Table S8 GO term enrichment).

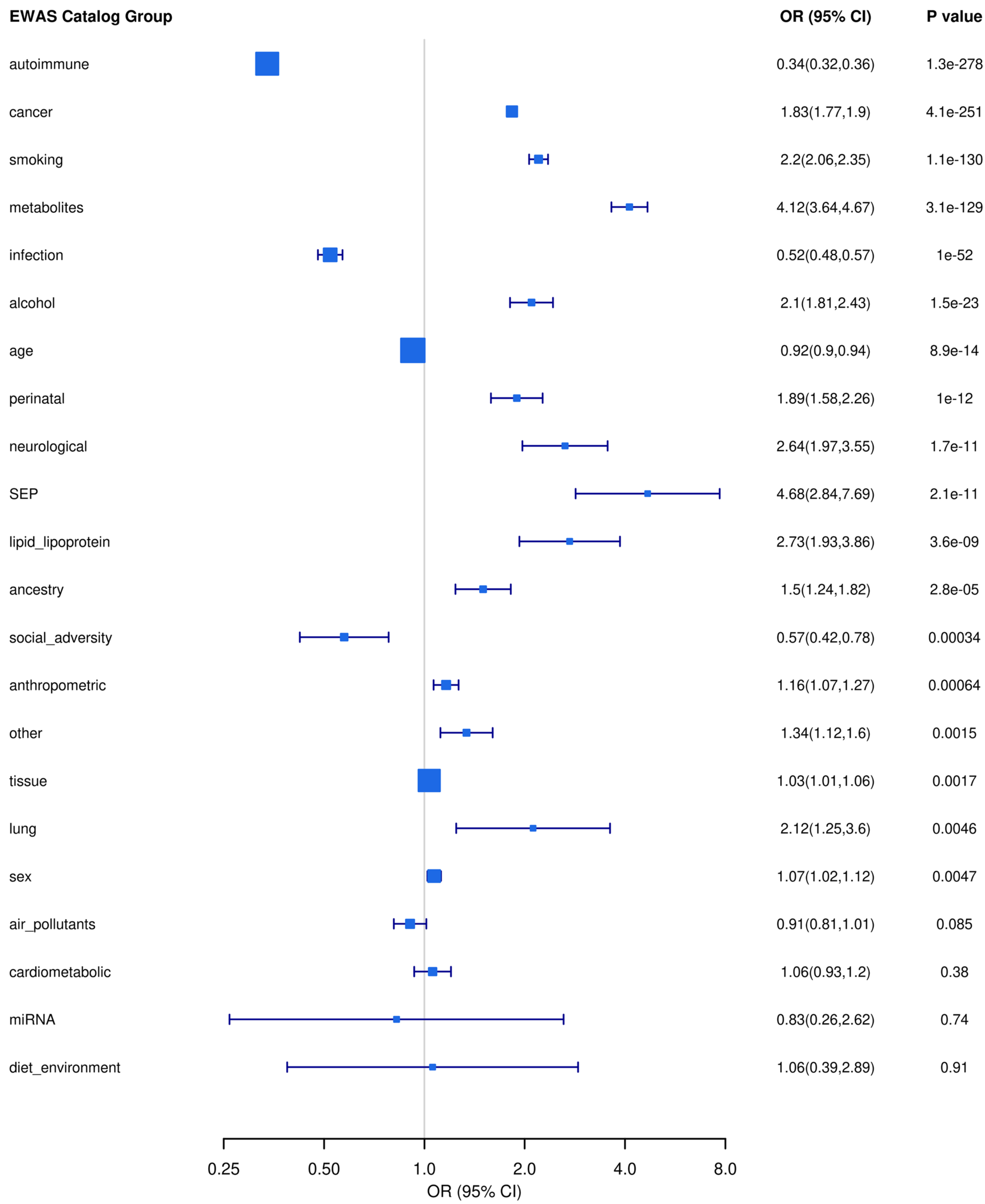

Fig. 3 Enrichment of EWAS Catalog phenotypes amongst ethnicity-associated CpG sites. Entries from the EWAS Catalog were reduced into categories of related phenotypes [56]. Group "other" contained CpGs not assigned categories and represented 0.24% of all unique CpGs across categories. *SEP* Socio-economic position

Regulatory pathways overlap with the genomic position of ethnicity-associated CpGs

Using LOLA, we identified regulatory elements overlapping with the genomic position of lower and higher methylation sets associated with ethnicity (Additional file 4: Table S9). Overlaps in the higher methylation set include sites annotated to CpG islands, DNAse hypersensitivity sites across multiple tissues including stem, brain, liver and hematopoietic cells and transcription factor binding sites (most strongly EZH2, a component of the polycomb repressive complex 2 (PRC2) which functions to methylate Lys-27 on histone 3). The higher methylation set also includes overlaps with histone modifications with the strongest evidence for overlaps with H3K4me1 and H3K27me3 which indicate active enhancer regions [58]. This is also supported by overlaps with enhancer segments across multiple cell types.

In the lower methylation set, the overlaps were fewer than in the higher methylation set ($n=74$ vs $n=299$). Overlaps in the lower methylation set include those with repressed segments across multiple cell types and UCSC repeat regions. DNAse hypersensitivity site overlaps in the lower methylation set were observed predominantly in hematopoietic cells. Transcription factor binding site overlaps include c-Fos and C-Jun which combine to form activator protein-1, a transcription factor which coordinates transcription in response to cytokines and infection by viruses and bacteria [59]. In contrast to the higher methylation set, there were no overlaps with histone modifications in the lower methylation set (see Additional file 4: Table S9).

Discussion

We observed genome-wide differences in DNA methylation between South Asian and European individuals at 16,344 CpG sites and regions with effect sizes that were comparable in the magnitude of effect to those seen in EWAS of smoking [29, 31, 51]. These differences were replicated between two distinct cohorts: one of men in older adulthood and one of women in pregnancy. By using DNA samples given at follow-up clinics in SABRE, we showed temporal stability of ethnic differences in DNA methylation over a c.20-year period. We also observed lower levels of systemic inflammation (measured by mdNLR) in South Asians compared to Europeans in both cohorts. Ethnic differences in epigenetic age acceleration were found for some of the estimators tested (namely PhenoAge and SkinBlood), but only in SABRE. Functional exploration of the ethnic differences observed indicated that ethnicity-associated CpG sites are depleted in 3'UTRs and intronic regions and highly enriched for eQTMs. By comparing ethnicity-associated CpG sites with EWAS Catalog entries, we identified that ethnicity-associated CpGs are enriched amongst multiple groups of traits including metabolites and autoimmune phenotypes that had previously been reported to be associated with variation in DNA methylation. On further investigation within the SABRE cohort, we found that ethnic differences are predominantly explained by differences in cell composition, and to a lesser extent explained by smoking and genetic effects.

Initial exploration of genetic and DNA methylation variance using PCA demonstrated that individuals from each self-reported ethnic group were genetically and epigenetically distinct. Genetic variation estimated from genetic PCA is greater in South Asian compared to European individuals in SABRE and BiB. This is expected since South Asian individuals in these two cohorts have a recent migration history from a large geographical area into the UK. Methylation PCA analysis also separated self-reported ethnicities but this was less distinct than for genetic data.

Ethnicity associations in SABRE were temporally stable (effect estimates $R^2=0.9$ over 20 years) although we acknowledge there may be a healthy survivor effect since around 25% of the cohort died between baseline and follow-up clinics. In previous studies, CpG sites stable over time have been postulated to be driven by genetic effects [60]. However, our data suggest that genetic variation is not a strong determinant of ethnicity-associated differences in DNA methylation. We therefore suggest that the DNA methylation differences observed are likely to be predominantly environmentally determined and persist across the life course. For example, this hypothesis is consistent with reported observations of early life exposures leaving a lasting impact on the methylome [61].

A number of studies have identified differences in epigenetic age acceleration between ancestries [62–64], although none has previously studied South Asian individuals. Self-reported South Asian individuals had higher age acceleration amongst all epigenetic clock measures than those in the European ethnic group in SABRE. For Levine PhenoAge and Horvath SkinBlood estimators, these differences were robust to adjustment for BMI and self-reported smoking. We did not replicate results in BiB but postulate this may be due to the younger age of BiB participants. It has been suggested that epigenetic clocks may contain ancestry-specific CpGs in their models so further exploration of clock models is required before concrete inferences can be made with respect to the biological or clinical significance of our observations [65].

An index of systemic inflammation derived from relative proportions of white blood cell types (mdNLR) was substantially lower in South Asians compared to Europeans in both SABRE and BiB and the association was not explained by age, smoking or BMI. The presence of differences between ethnicities in this cell count-derived measure accords with observations in this study of cell counts being the major determinant of ethnic differences in DNA methylation, as discussed below. Higher mdNLR has previously been associated with increased cancer [66–68], rheumatoid arthritis [69] and cardiovascular disease risk [70]. In previous studies, differences in mdNLR by ethnicity have been reported with European ancestry groups having the highest level of mdNLR [50, 71]. Our results include South Asian ancestries for the first time and support Europeans having elevated mdNLR compared to other ancestries. However, this result may represent higher systemic inflammation in Europeans compared to South Asians or may represent a difference in reference ranges between groups. It may also reflect sample selection within our cohorts, for example, the SABRE cohort was limited to individuals free from diabetes or coronary heart disease. In comparison with disease-free individuals, South Asians are typically healthier than European individuals, for example, reporting lower levels of smoking (see Table 1).

Exploration of EWAS results identified that ethnic differences in cell composition were the major driver of ethnic differences in DNA methylation between South Asians and Europeans. This finding aligns with other research highlighting cellular composition as a major source of variation in DNA methylation measured in blood [72]. This highlights the importance of accounting for cell composition in epigenetic studies, especially when there may be the heterogeneity of genetic ancestries within a study.

We investigated the contribution of common genetic (SNP) effects to observed differences in DNA methylation between ethnic groups. In line with other studies [53], we identify that mQTL effects are common but small in magnitude. In addition, we identified 67 interaction effects between mQTLs and ethnicity indicating there may be associations between methylation and ethnicity that are attributable to ethnic-specific genetic effects. These findings are of interest and require further exploration particularly given the mQTL reference panel consisted of European ancestry individuals. We may not have therefore captured SNP effects specific to individuals of South Asian ethnicity within SABRE.

In addition to exploring the potential determinants of DNA methylation differences between Europeans and South Asian ancestral groups, we sought to investigate whether there was any evidence to support the hypothesis that these ethnicity-variable CpG sites may be associated with biological functions (using GO, eQTM and LOLA) or disease phenotypes or risk factors (using the EWAS Catalog). We acknowledged earlier that European and South Asian ethnic groups show marked discordance in their risk for some non-communicable diseases, and we postulate that variation in DNA methylation may contribute to this. We show that ethnicity-associated CpG sites are not randomly distributed throughout the genome and are more likely to be eQTMs than chance alone. Using LOLA, we identified overlaps between regulatory elements and CpG sites associated with ethnicity suggesting ethnicity-associated CpG sites were enriched in active enhancer regions. These findings suggest that ethnicity-associated CpG sites are enriched for functionally important sites.

Comparing ethnicity-associated CpG sites with those in the EWAS Catalog [42], we identified enrichment for CpG sites associated with ancestry. This gives further support to our EWAS results. We also observed enrichment amongst CpG sites associated with sex and smoking. Since our analyses are stratified by sex and adjusted for smoking this may suggest residual confounding or highlight that other studies reporting their associations in the EWAS Catalog may have not adjusted for these covariates. We observed strong enrichment amongst CpG sites associated with disease phenotypes, most prominently an enrichment of ethnicity-associated CpG sites associated with metabolites and a depletion of ethnicity-associated CpG sites amongst CpG sites associated with autoimmune diseases amongst others. This finding supports an avenue of further research to define ethnic-specific risk. However, SABRE samples are restricted to healthy individuals (e.g. with no existing diabetes or cardiovascular disease) so this limits the conclusions we can make on phenotype enrichment compared to the EWAS Catalog data.

Strengths and limitations

This is one of the largest studies to explore differences across the epigenome between South Asian and European populations and makes an important contribution to an emerging but still small literature on ethnic differences in epigenetic markers. We have also replicated epigenome-wide ethnic differences in an independent cohort (BiB). That the replication sample (women only with DNA methylation measured on pregnancy samples) was very different to the discovery sample (men only in older age) has some advantage, in that where replication occurs this is likely to reflect robust ethnic differences. Additional strengths include the availability of genomic and other data in the two studies that enabled the exploration of factors relating to ethnic differences in DNA methylation.

We acknowledge some key limitations. All reference data utilised is predominantly from data sets of European ancestry. For example, mQTLs from GoDMC were identified in individuals of European ancestry [53]. Therefore, we may not have identified mQTLs specific to individuals of South Asian ancestry which would mean that genetic effects may be underestimated in our study. EWAS studies in the EWAS Catalog are also predominantly conducted in European ancestry individuals. There may also be bias in methylation-derived variables (cell composition and mdNLR) which rely on a European ancestry reference set [49, 50]. The main EWAS was analysed in male participants and replication in females. There may therefore be sex-specific ethnic differences that we could not detect in this analysis. In addition, both SABRE and BiB samples are restricted to selected subsets of healthy individuals (see methods). This potentially introduces selection bias [73] and limits the conclusions we can make on phenotype enrichment compared to the EWAS Catalog data. Finally, individuals in SABRE and BiB were recruited in the UK, with South Asian individuals representing migrant populations. This may limit generalisability to non-migrant South Asians vs Europeans.

This study has only made comparisons between two self-reported ethnic groups and models a limited number of phenotypes. Wider efforts are required to increase population diversity in epigenetic studies.

Conclusions

This study aimed to define and characterise differences in DNA methylation between individuals of European and South Asian self-reported ethnicity. This is important research because there are little epigenetic data collected across global populations and very few cross-ancestry comparisons made. This study highlights widespread differences in DNA methylation throughout the genome between South Asians and Europeans. We estimate that 76% of ethnic differences observed in this study are attributable to differences in cell composition. Surprisingly, little ethnicity-associated variation in DNA methylation was explained by underlying genetic differences between the groups. We also identified that South Asians have lower levels of systemic inflammation (using the methylation-derived neutrophil-to-lymphocyte ratio), but in the older SABRE cohort we also observed higher age acceleration in South Asian individuals compared to Europeans. We used public databases to characterise ethnicity-associated CpG sites and show they are enriched for eQTMs and GO terms, depleted in 3'UTRs and intronic regions and have overlaps with regulatory markers. This study demonstrates that epigenetic differences between ethnicities are widespread. It highlights the need to consider cell composition, genetic variation and lifestyle confounders in population studies, particularly in studies with participants from different ancestries. This study also identifies that there are many CpG sites that are associated with methylation independently of these factors. Further exploration of these ethnicity-associated CpG sites could improve our understanding of disease aetiology and refine predictive models which rely on epigenetic data.

Methods

Study populations

SABRE

Samples were derived from the extensively characterised population-based Southall And Brent REvisited (SABRE) cohort [3]. The SABRE cohort includes 1711 first-generation South Asian migrants and 1762 European-origin individuals aged between 40 and 69 living in west London, UK. Initial investigations were carried out between 1988 and 1991 with follow-up clinics c.20 years later.

DNA was extracted from peripheral blood samples collected at baseline and follow-up visits. A random sample of 800 men from the SABRE cohort at baseline clinic was selected for this study. Sampling was limited to individuals free from diabetes or coronary heart disease with good-quality DNA available at baseline. Samples were stratified across two age groups (cut at the median of 52 years) and across two self-reported ethnic groups (South Asian and European). Samples were randomly selected to give equal numbers of samples in each stratum. Repeat DNA was collected at follow-up clinics where possible ($n = 139/800$ individuals).

The study was approved by St Mary's Hospital Research Ethics Committee (07/H0712/109) and all participants provided written informed consent.

Born in Bradford

Born in Bradford (BiB) is a multi-ethnic pregnancy and birth cohort that recruited pregnant women largely during an oral glucose tolerance test (24–28 weeks of gestation) in Bradford, UK, between 2007 and 2010 [43]. The cohort has been described in detail elsewhere [43, 44, 74]. In total 12,453 women, who gave birth to 13,818 infants were recruited. DNA was extracted from peripheral blood samples. DNA methylation was generated from a non-random subsample of 1000 BiB maternal DNA samples, that prioritised the two largest (self-reported) ethnic groups (White British and Pakistani) and included women who had a singleton pregnancy and where both mother and their child had genome-wide data. Samples were selected to give equal numbers of self-reported white British ($n = 444$) and Pakistani ($n = 472$) mother–offspring pairs.

Ethical approval for the data collection was granted by Bradford Research Ethics Committee (Ref 07/H1302/112). All participants gave written informed consent.

Methylation data

DNA from SABRE participants was analysed using HumanMethylation450 BeadChips (Illumina, San Diego, CA, USA). Approximately 500 ng of DNA was bisulphite modified using EZ DNA methylation kits (Zymo Research, Orange, CA, USA). The manufacturer's protocol was followed using the alternative incubation conditions recommended when using Illumina BeadChips. BeadChips were processed according to the manufacturer's protocol and Illumina iScan software v3.3.28 was used to scan the arrays. Quality control and processing were conducted according to the pipeline described previously [75] but with more stringent quality control parameters using *meffil.qc.parameters()*: The detection p value threshold was reduced to 0.1 for samples and probes, bead number threshold was reduced to 0.1 for samples and probes, sample genotype concordance was reduced to 0.8 and sex outliers < 5 SD were removed. In total, 11 samples and 1644 probes were removed from the data set. SABRE data were normalised using 15 control probe PCs derived from the technical probes informed by *meffil* scree plots [75]. During normalisation of raw data, batch effects (defined as 450 k slide ID) were removed where the slide was modelled as a random effect. Cell composition was estimated from methylation data using the Houseman method [49, 76]. Statistical analyses were performed on β values throughout. Samples with missing genetic data or admixture (Additional file 1: Fig. 1) were also removed leaving 747 samples and 484,781 probes available for analysis.

In BiB, the data generation pipeline was analogous to that of SABRE, but the methylation array used was the HumanMethylationEPIC BeadChip (Illumina, San Diego, CA, USA). Following QC steps, 922 samples and 860,750 probes were available for analysis.

Methylation principal components

CpG sites were pruned by ranking on variance and removing CpGs with correlation $R^2 > 0.2$. In each correlated pair the CpG with the highest variance was retained. This pruning resulted in a reduced data set of 21,023 CpGs (SABRE) and 19,438 CpGs (BiB). Principal component analysis (PCA) then was performed to generate methylation principal components using *prcomp*.

Genetic data

DNA from 2980 SABRE participants were genotyped using the UCL druggable target array, comprising the Illumina Human 1679 Core Bead Chip (~ 240 K genome-wide markers) and an additional custom set of 200 K markers on genes encoding proteins involved in drug handling, drug action, and druggable targets. This was developed in collaboration with the London School of Hygiene and Tropical Medicine and the European Bioinformatics Institute.

For both the SABRE South Asian and European sub-cohorts, individuals were excluded on the basis of incorrect sex assignment, high missingness (> 5%), abnormal heterozygosity (het > mean (het) + 3*sd (het) or het < mean (het) − 3*sd (het)), cryptic relatedness (pihat ≥ 0.9999 identified 20 cases of sample mislabelling which were excluded) and non-concordant ancestry (detected via Principal Components Analysis).

SABRE phasing was performed using Eagle and ethnicity-specific imputation was performed using Minimac3 against the HRC reference panel (http://www.haplotype-reference-consortium.org/) [77].

SABRE genotypes were filtered to have a MAF > 0.01, imputation info score > 0.8 and HWE $p < 0.00001$. After data cleaning and QC, 1527 participants of European ethnicity and 1210 participants of South Asian ethnicity had available 6,912,559 and 6,046,044 SNPs, respectively.

For BiB participants, DNA was genotyped using either Infinium Human Core Exome-24 v1.1 arrays or Infinium global screen-24 + v1.0 arrays (Illumina, San Diego, CA, USA). Samples were pre-processed using GenomeStudio 2011.1. Samples with Call Rate < 0.95 were excluded. Poorly performing SNPs were identified based on Call Freq < 0.97, Cluster Sep ≤ 0.3, AB R Mean ≤ 0.2, BB R Mean ≤ 0.2, AA R Mean ≤ 0.2, 10% GC Score ≤ 0.2, MI Errors < 2 and Rep Errors < 2.

After quality control, a VCF containing 15,628 BiB participants and 459,340 genomic variants was submitted to the Sanger Imputation Service using the "UK10K + 1000 Genomes Phase 3" as a reference panel and "pre-phase with EAGLE2 and impute" as the pipeline. The 1000 Genomes Phase 3 panel was chosen as it contains samples originating from several global populations [78]. The resultant imputation data set contains 15,628 participants and 87,558,135 variants.

Genetic principal components

For the SABRE cohort, we combined European and South Asian populations on intersecting imputed SNPs ($n = 5{,}539{,}760$) and the availability of DNA methylation data. In total, 749 European and South Asian individuals were included in the genetic principal component analysis.

For the BiB cohort, there were 875 individuals after sub-setting for participants with available DNA methylation and genetic data.

The HapMap Phase 3 (HapMap3) reference data set contains around 1.5 million SNPs genotyped in 1397 individuals from a variety of populations. This data set is available from the HapMap FTP site (ftp://ftp.ncbi.nlm.nih.gov/hapmap/). We used LiftOver to convert HapMap3 genomic positions from Build 36 to 37 (hg18 to hg19 chain file) and merged with the SABRE and BiB cohort independently on intersecting SNPs ($N=1{,}032{,}847$ (SABRE) and $N=1{,}348{,}525$ (BiB)) giving a final sample of 2146 individuals for SABRE and HapMap3 combined, and 2272 individuals for BiB and HapMap3 combined.

Principal component analysis (PCA) was performed to generate the first 20 genetic principal components (PCs) for i) the SABRE cohort and BiB cohorts (independently) and ii) each cohort merged with HapMap3 using Plink 1.90 [79]. There were 1,032,847 SNPs intersecting between SABRE and HapMap3 and 1,348,525 SNPs intersecting between BiB and HapMap3. After pruning for independent SNPs (window size = 50 SNPs, step size = 5 SNPs, VIF = 1.5, and excluding long-range LD regions); 103,973 and 120,671 SNPs were taken forwards for PCA in the SABRE cohort and combined SABRE and HapMap3, respectively. For BiB, there were 212,061 and 252,916 SNPs taken forwards for PCA in the BiB cohort and combined BiB and HapMap3, respectively.

Derived phenotypes

Measures of epigenetic age and age acceleration were generated using the *methylclock* R package [48]. The following estimators were utilised: Horvath, Hannum, Alfonso and Gonzalez, Horvath Skin and Blood, Levine PhenoAge [80–85]. Extrinsic Epigenetic Age accelerations (EEAAs) for each method were calculated as the residuals from a linear model where DNAmAge was the outcome and chronological age was the predictor. Intrinsic epigenetic age accelerations (IEAA) were calculated as the residuals from a linear model, where DNAmAge was used as the outcome and chronological age and cell counts were predictors [36].

The methylation-derived neutrophil-to-lymphocyte ratio (mdNLR) is an epigenetic estimate of neutrophil-to-lymphocyte ratio [50]. Using cell count estimates, the neutrophil count was divided by the lymphocyte cell count (calculated as the sum of B cells, CD4+T cells, CD8+T cells and natural killer cells) to provide the mdNLR.

Statistical analysis

Baseline characteristics comparing South Asians and Europeans were conducted using *t* tests for continuous and chi-squared tests for categorical variables.

EWAS were conducted utilising *meffil,* where methylation was the outcome and ethnicity was the predictor. Covariates in each model are described in the results. Differentially methylated regions (DMRs) were identified by combining EWAS test statistics between consecutive CpG sites (maximum distance between sites = 500 base pairs) using the R package *dmrff* [86].

To replicate our findings in an alternative cohort, we conducted an EWAS and DMR analysis of self-reported ethnicity amongst individuals from the Born in Bradford cohort.

To investigate inflation of observed λ values we used a method described previously [45] where lambdas were recalculated in SABRE using CpGs which had $p > 0.2$ in the corresponding BiB EWAS and vice versa where lambdas were recalculated in BiB using CpGs which had $p > 0.2$ in the corresponding SABRE EWAS.

For epigenetic age analyses, linear models were used to identify differences in age acceleration between SABRE and BiB ethnic groups where ethnicity was the predictor and age acceleration was the outcome. In adjusted models, BMI and smoking were included as additional covariates in the models as these are common disease risk factors and differed by ethnicity in SABRE and BiB.

For mdNLR analyses, linear models were used to identify differences in mdNLR between SABRE and BiB ethnic groups where ethnicity was the predictor and mdNLR was the outcome. In adjusted models, BMI, smoking and age were included as additional covariates in the models as they are possible mediators in the relationship between ethnicity and mdNLR.

The Genetics of DNA Methylation Consortium (GoDMC) database (http://mqtldb.godmc.org.uk/) [53] was used to identify mQTLs of ethnicity-associated CpG sites identified in SABRE. The mQTLs retrieved were restricted to those at $p < 1 \times 10{-}8$ (*cis* mQTLs) or $p < 1 \times 10^{-14}$ (*trans*-mQTLs) in accordance with the GoDMC meta-analysis protocol [53]. For SNP analyses, allele frequencies were compared between populations using Fisher's tests. In regression models, SNPs were coded as additive effects.

Genetic associations with methylation and gene–environment interactions were modelled using *GEM*, using self-reported ethnicity as the environmental component [54].

Genomic region information was annotated through ANNOVAR [87]. CpG sites were annotated into functional categories including downstream, exonic, exonic/splicing, intergenic, intronic, ncRNA_exonic, ncRNA_exonic/splicing, ncRNA_intronic, ncRNA_splicing, splicing, upstream, upstream/downstream, UTR3, UTR5 and UTR5/UTR3. Enrichment was assessed using chi-squared tests.

Ethnicity-associated CpG sites were compared to those listed in EWAS Catalog (http://www.ewascatalog.org/) in June 2021. The EWAS Catalog contains EWAS studies analysing at least 100,000 CpG sites using a minimum sample size of 100 individuals. Catalog entries were reduced into related categories [56], and enrichment was assessed using Wald odds ratio and chi-squared tests.

We assessed whether ethnicity-associated CpG sites were also *cis*-expression quantitative trait methylation (*cis*-eQTMs). *Cis*-eQTM data were extracted from BIOS consortium analyses of methylation and gene expression from blood samples in 2101 Dutch individuals [57]. Enrichment was assessed using Wald odds ratio and chi-squared tests.

Biological enrichment analyses were run using unique CpG sites identified from the EWAS and DMR analyses as input. Since increased and decreased levels of methylation at CpG sites in our analysis were likely to be biologically distinct, we stratified our enrichment analyses accordingly. This approach has been used previously [88]. Gene ontology enrichment [89, 90] was conducted using *missMethyl* R package [91]. Further enrichment analyses were conducted using the *LOLA* R package [92]. LOLA input was the genomic coordinates of ethnicity-associated CpG sites and the background set was the genomic coordinates of all array CpG sites included in the EWAS. The LOLA core region set was used to test for enrichments. For each LOLA analysis, results were filtered to retain enriched region sets where the support (i.e. number of regions overlapping) ≥ 5 and the q value was < 0.05.

Analyses were conducted in R, version 4.1.2 (http://www.r-project.org).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13148-022-01351-2.

Additional file 1. Figure S1. Principal component analysis of SABRE genetic data. The upper panels show PCs 1 and 2 generated from SABRE or BiB data. The lower panels show PCs 1 and 2 generated from SABRE + HapMap3 data or BiB + HapMap3 data. Colours indicate self-reported ethnic group or subgroup (SABRE, BiB) or population group (HapMap3). Axis labels show the variance explained by each PC. HapMap3 populations have been collapsed: South Asian = GIH; European = CEU + TSI; African = ASW + LWK + MKK; Mexican = MEX; South East Asian = CHB + CHD + JPT. In the lower panel (SABRE), two outliers self-reporting as South Asian in SABRE data appear intermediate between African groups from HapMap3 and the remaining South Asian cluster. Both of these individuals reported their country of birth as an African country indicating possible genetic admixture in these individuals. These two individuals were removed from all other analyses. In the lower panel (SABRE), "other South Asian" individuals predominantly identify their country of birth as South Asia (India, Pakistan, Bangladesh, Sri Lanka), $n = 42/50$.

Additional file 2. Figure S2. Principal component analysis of SABRE DNA methylation data. The upper panel shows PCs 1 and 2 generated from the 1000 most variant probes in SABRE or BiB. The lower panel shows PCs 1 and 2 generated from the 10000 most variant probes in SABRE or BiB. Colours indicate self-reported ethnic subgrouping (SABRE, BiB). Axis labels show the variance explained by each PC.

Additional file 3. Table S1. EWAS results.

Additional file 4. Table S2. DMR results. **Table S3.** Epigenetic age acceleration results. **Table S4.** Cell count differences. **Table S5.** Effects of adjustment for mQTL and allele frequency differences. **Table S6.** Effects of adjustment for mQTL and allele frequency differences. **Table S7.** ANNOVAR enrichment. **Table S8.** GO term enrichment. **Table S9.** LOLA enrichment.

Additional file 5. Figure S3. Comparison of effect sizes for cell composition adjustment of EWAS. Each CpG is represented by a point on the graph with 95% confidence intervals for effect estimates. Red dashed line: linear regression between data sets. Black dashed line: line of equality. Orange highlighted estimates: $p \leq 1.03 \times 10^{-7}$ cell adjusted EWAS ($n = 3922/16{,}344$ CpG sites).

Acknowledgements

We gratefully acknowledge the participation of all individuals contributing to SABRE and Born in Bradford cohorts. Born in Bradford is only possible because of the enthusiasm and commitment of the Children and Parents in BiB. We are grateful to all the participants, teachers, school staff, health professionals and researchers who have made Born in Bradford happen.

Author contributions

HRE conceptualised the study with input from CLR, JLM and KB. HRE led the data analysis and wrote the manuscript. KB conducted genetic PCA analysis. TT, NC and ADH contributed to SABRE data. GDS, CLR, NC and ADH contributed to project funding. DAL, DM, GS and JW contributed to the recruitment, data collection and management of BiB data used here. DAL and JW obtained funds for the BiB data used here. All authors contributed to the discussion, provided critical input to the project and drafted the manuscript. All authors read and approved the final manuscript.

Funding

The SABRE study was funded at baseline by the Medical Research Council, Diabetes UK, and the British Heart Foundation. At follow-up, the study was funded by the Wellcome Trust and the British Heart Foundation. Methylation analysis in the SABRE cohort was supported by a Wellcome Trust Enhancement grant 082464/Z/07/C. Genotyping analysis in the SABRE cohort was supported by the British Heart Foundation (CS/13/1/30327). BiB receives core infrastructure funding from the Wellcome Trust (WT101597MA), a joint grant from the UK Medical Research Council (MRC) and Economic and Social Science Research Council (ESRC) (MR/N024397/1), the British Heart Foundation (CS/16/4/32482) and the National Institute for Health Research (NIHR) under its Applied Research Collaboration (ARC) for Yorkshire and Humber and the Clinical Research Network (CRN). DNA methylation data was funded by the UK Medical Research Council via the Integrative Epidemiology Unit (MC_UU_12013/5). HRE, KB, JLM, CLR, DAL and GDS work in the Medical Research Council Integrative Epidemiology Unit at the University of Bristol, which is supported by the Medical Research Council and the University of Bristol (MC_UU_00011/5, MC_UU_00011/6 and MC_UU_00011/1). DAL is supported by a British Heart Foundation Chair (CH/F/20/90003) and National Institute for Health Research Senior Investigator award (NF-0616-10102). NC and AH received support from a Biomedical Research Centre Award to Imperial NHS Healthcare Trust. The funders played no role in the study design and conduct, or in these analyses or the decision to submit the manuscript for publication. The SABRE study group is entirely independent of the funding bodies.

Availability of data and materials

SABRE data used for this submission will be made available on request to mrclha.swiftinfo@ucl.ac.uk. Further details regarding data sharing can be found on the cohort web pages (https://www.sabrestudy.org/home-2/data-sharing/). Born in Bradford data used for this submission will be made available on request to the Born in Bradford executive committee (borninbradford@bthft.nhs.uk). The Born in Bradford data management plan (available here: https://borninbradford.nhs.uk/research/how-to-access-data/) describes in detail the policy regarding data sharing, which is through a system of managed open access.

Declarations

Ethics approval and consent to participate

The SABRE study was approved by St Mary's Hospital Research Ethics Committee (07/H0712/109), and all participants provided written informed consent. Ethical approval for the data collection in BiB was granted by Bradford

Research Ethics Committee (Ref 07/H1302/112). All participants gave written informed consent.

Consent for publication
Not applicable.

Competing interests
DAL has received support from Roche Diagnostics and Medtronic Ltd. during the last 10 years for research unrelated to this study. NC serves on and receives remuneration for data monitoring and safety committees of trials sponsored by AstraZeneca. Other authors declare no conflict of interest.

Author details
[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK. [2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK. [3]Department of Population Science and Experimental Medicine, Institute of Cardiovascular Science, University College London, London, UK. [4]MRC Unit for Lifelong Health and Ageing, University College London, London, UK. [5]Bradford Institute for Health Research, Bradford, UK.

Received: 8 June 2022 Accepted: 20 September 2022
Published online: 15 October 2022

References

1. Agyemang C, van den Born BJ. Non-communicable diseases in migrants: an expert review. J Travel Med. 2019. https://doi.org/10.1093/jtm/tay107.
2. Ali R, Chowdhury A, Farouhi N, Wareham N. Ethnic disparities in the major causes of mortality and their risk factors in the UK – submission to the Commission on Race and Ethnic Disparities. 2021. https://www.gov.uk/government/publications/the-report-of-the-commission-on-race-and-ethnic-disparities-supporting-research/ethnicdisparities-in-the-major-causes-of-mortality-and-their-risk-factors-by-dr-raghib-ali-et-al.
3. Tillin T, Forouhi NG, McKeigue PM, Chaturvedi N. Southall And Brent REvisited: Cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins. Int J Epidemiol. 2012;41:33–42. https://doi.org/10.1093/ije/dyq175.
4. Sattar N, Gill JM. Type 2 diabetes in migrant south Asians: mechanisms, mitigation, and management. Lancet Diabetes Endocrinol. 2015;3:1004–16. https://doi.org/10.1016/S2213-8587(15)00326-5.
5. Arnold M, Razum O, Coebergh JW. Cancer risk diversity in non-western migrants to Europe: an overview of the literature. Eur J Cancer. 2010;46:2647–59. https://doi.org/10.1016/j.ejca.2010.07.050.
6. Tillin T, et al. Insulin resistance and truncal obesity as important determinants of the greater incidence of diabetes in Indian Asians and African Caribbeans compared with Europeans: the Southall And Brent REvisited (SABRE) cohort. Diabetes Care. 2013;36:383–93. https://doi.org/10.2337/dc12-0544.
7. Tillin T, et al. The relationship between metabolic risk factors and incident cardiovascular disease in Europeans, South Asians, and African Caribbeans: SABRE (Southall and Brent Revisited)—a prospective population-based study. J Am Coll Cardiol. 2013;61:1777–86. https://doi.org/10.1016/j.jacc.2012.12.046.
8. Tillin T, et al. Ethnicity-specific obesity cut-points in the development of Type 2 diabetes—a prospective study including three ethnic groups in the United Kingdom. Diabet Med. 2015;32:226–34. https://doi.org/10.1111/dme.12576.
9. Nyamdorj R, et al. Ethnic comparison of the association of undiagnosed diabetes with obesity. Int J Obes (Lond). 2010;34:332–9. https://doi.org/10.1038/ijo.2009.225.
10. Williams ED, Nazroo JY, Kooner JS, Steptoe A. Subgroup differences in psychosocial factors relating to coronary heart disease in the UK South Asian population. J Psychosom Res. 2010;69:379–87. https://doi.org/10.1016/j.jpsychores.2010.03.015.
11. Williams R, Bhopal R, Hunt K. Coronary risk in a British Punjabi population: comparative profile of non-biochemical factors. Int J Epidemiol. 1994;23:28–37. https://doi.org/10.1093/ije/23.1.28.
12. Kapadia D et al. (ed) NHS Race and Health Observatory; 2022.
13. Sun YQ, et al. Assessing the role of genome-wide DNA methylation between smoking and risk of lung cancer using repeated measurements: the HUNT study. Int J Epidemiol. 2021. https://doi.org/10.1093/ije/dyab044.
14. Michaud DS, et al. Epigenome-wide association study using prediagnostic bloods identifies new genomic regions associated with pancreatic cancer risk. JNCI Cancer Spectr. 2020;4:pkaa041. https://doi.org/10.1093/jncics/pkaa041.
15. Xu Z, Sandler DP, Taylor JA. Blood DNA methylation and breast cancer: a prospective case-cohort analysis in the sister study. J Natl Cancer Inst. 2020;112:87–94. https://doi.org/10.1093/jnci/djz065.
16. Koestler DC, et al. Distinct patterns of DNA methylation in conventional adenomas involving the right and left colon. Mod Pathol. 2014;27:145–55. https://doi.org/10.1038/modpathol.2013.104.
17. Davegardh C, Garcia-Calzon S, Bacos K, Ling C. DNA methylation in the pathogenesis of type 2 diabetes in humans. Mol Metab. 2018;14:12–25. https://doi.org/10.1016/j.molmet.2018.01.022.
18. Juvinao-Quintero DL, et al. DNA methylation of blood cells is associated with prevalent type 2 diabetes in a meta-analysis of four European cohorts. Clin Epigenet. 2021;13:40. https://doi.org/10.1186/s13148-021-01027-3.
19. Florath I, et al. Type 2 diabetes and leucocyte DNA methylation: an epigenome-wide association study in over 1,500 older adults. Diabetologia. 2016;59:130–8. https://doi.org/10.1007/s00125-015-3773-7.
20. Liu Y, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31:142–7. https://doi.org/10.1038/nbt.2487.
21. Hannon E, et al. DNA methylation meta-analysis reveals cellular alterations in psychosis and markers of treatment-resistant schizophrenia. Elife. 2021. https://doi.org/10.7554/eLife.58430.
22. Huang Y, et al. Identification, heritability, and relation with gene expression of novel DNA methylation loci for blood pressure. Hypertension. 2020;76:195–205. https://doi.org/10.1161/HYPERTENSIONAHA.120.14973.
23. Richard MA, et al. DNA methylation analysis identifies loci for blood pressure regulation. Am J Hum Genet. 2017;101:888–902. https://doi.org/10.1016/j.ajhg.2017.09.028.
24. Gomez-Alonso MDC, et al. DNA methylation and lipid metabolism: an EWAS of 226 metabolic measures. Clin Epigenet. 2021;13:7. https://doi.org/10.1186/s13148-020-00957-8.
25. Wahl 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.
26. Aslibekyan S, et al. Epigenome-wide study identifies novel methylation loci associated with body mass index and waist circumference. Obesity (Silver Spring). 2015;23:1493–501. https://doi.org/10.1002/oby.21111.
27. Demerath EW, 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;24:4464–79. https://doi.org/10.1093/hmg/ddv161.
28. Geurts YM, et al. Novel associations between blood DNA methylation and body mass index in middle-aged and older adults. Int J Obes (Lond). 2018;42:887–96. https://doi.org/10.1038/ijo.2017.269.
29. Christiansen C, et al. Novel DNA methylation signatures of tobacco smoking with trans-ethnic effects. Clin Epigenet. 2021;13:36. https://doi.org/10.1186/s13148-021-01018-4.
30. Domingo-Relloso A, et al. Cadmium, smoking, and human blood DNA methylation profiles in adults from the strong heart study. Environ Health Perspect. 2020;128:67005. https://doi.org/10.1289/EHP6345.
31. Joehanes R, et al. Epigenetic signatures of cigarette smoking. Circ Cardiovasc Genet. 2016;9:436–47. https://doi.org/10.1161/CIRCGENETICS.116.001506.
32. Maas SCE, et al. Validated inference of smoking habits from blood with a finite DNA methylation marker set. Eur J Epidemiol. 2019;34:1055–74. https://doi.org/10.1007/s10654-019-00555-w.
33. Liu C, et al. A DNA methylation biomarker of alcohol consumption. Mol Psychiatry. 2018;23:422–33. https://doi.org/10.1038/mp.2016.192.
34. Dugue PA, et al. Alcohol consumption is associated with widespread changes in blood DNA methylation: analysis of cross-sectional and longitudinal data. Addict Biol. 2021;26:e12855. https://doi.org/10.1111/adb.12855.

35. Agha G, et al. Blood leukocyte DNA methylation predicts risk of future myocardial infarction and coronary heart disease. Circulation. 2019;140:645–57. https://doi.org/10.1161/CIRCULATIONAHA.118.039357.
36. Chen BH, et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. Aging (Albany NY). 2016;8:1844–65. https://doi.org/10.18632/aging.101020.
37. Marioni RE, et al. DNA methylation age of blood predicts all-cause mortality in later life. Genome Biol. 2015;16:25. https://doi.org/10.1186/s13059-015-0584-6.
38. Fraszczyk E, et al. Epigenome-wide association study of incident type 2 diabetes: a meta-analysis of five prospective European cohorts. Diabetologia. 2022. https://doi.org/10.1007/s00125-022-05652-2.
39. Galanter JM, et al. Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. Elife. 2017. https://doi.org/10.7554/eLife.20532.
40. Natri HM, et al. Genome-wide DNA methylation and gene expression patterns reflect genetic ancestry and environmental differences across the Indonesian archipelago. PLoS Genet. 2020;16: e1008749. https://doi.org/10.1371/journal.pgen.1008749.
41. Giri AK, et al. DNA methylation profiling reveals the presence of population-specific signatures correlating with phenotypic characteristics. Mol Genet Genom. 2017;292:655–62. https://doi.org/10.1007/s00438-017-1298-0.
42. Battram T, et al. The EWAS Catalog: a database of epigenome-wide association studies. Wellcome Open Res. 2022;7:41.
43. Raynor P, Born in Bradford Collaborative, G. Born in Bradford, a cohort study of babies born in Bradford, and their parents: protocol for the recruitment phase. BMC Public Health. 2008;8:327. https://doi.org/10.1186/1471-2458-8-327.
44. Wright J, et al. Cohort profile: the Born in Bradford multi-ethnic family cohort study. Int J Epidemiol. 2013;42:978–91. https://doi.org/10.1093/ije/dys112.
45. Solomon O, et al. Meta-analysis of epigenome-wide association studies in newborns and children show widespread sex differences in blood DNA methylation. Mutat Res Rev Mutat Res. 2022;789: 108415. https://doi.org/10.1016/j.mrrev.2022.108415.
46. Kent WJ, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006. https://doi.org/10.1101/gr.229102.
47. Rosenbloom KR, et al. ENCODE data in the UCSC genome browser: year 5 update. Nucleic Acids Res. 2013;41:D56-63. https://doi.org/10.1093/nar/gks1172.
48. Pelegi-Siso D, de Prado P, Ronkainen J, Bustamante M, Gonzalez JR. methylclock: a bioconductor package to estimate DNA methylation age. Bioinformatics. 2021;37:1759–60. https://doi.org/10.1093/bioinformatics/btaa825.
49. Houseman EA, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinform. 2012;13:86. https://doi.org/10.1186/1471-2105-13-86.
50. Koestler DC, et al. DNA methylation-derived neutrophil-to-lymphocyte ratio: an epigenetic tool to explore cancer inflammation and outcomes. Cancer Epidemiol Biomark Prev. 2017;26:328–38. https://doi.org/10.1158/1055-9965.EPI-16-0461.
51. Sikdar S, et al. Comparison of smoking-related DNA methylation between newborns from prenatal exposure and adults from personal smoking. Epigenomics. 2019;11:1487–500. https://doi.org/10.2217/epi-2019-0066.
52. Florath I, Butterbach K, Muller H, Bewerunge-Hudler M, Brenner H. Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. Hum Mol Genet. 2014;23:1186–201. https://doi.org/10.1093/hmg/ddt531.
53. Min JL, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. Nat Genet. 2021;53:1311–21. https://doi.org/10.1038/s41588-021-00923-x.
54. Pan H, Holbrook JD, Karnani N, Kwoh CK. Gene, Environment and Methylation (GEM): a tool suite to efficiently navigate large scale epigenome wide association studies and integrate genotype and interaction between genotype and environment. BMC Bioinform. 2016;17:299. https://doi.org/10.1186/s12859-016-1161-z.
55. Chen MH, et al. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. Cell. 2020;182:1198–213. https://doi.org/10.1016/j.cell.2020.06.045.
56. Elliott HR. Collapse EWAS catalog categories; 2021. https://github.com/hannah-e/collapse_EWAS_catalog_phenotypes/blob/9b65be66399d0c1d2fd71c2003dbf58e4e5b62ff/functional_analysis_regroup_EWAS_catalogue_phenotypes.R.
57. Bonder MJ, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017;49:131–8. https://doi.org/10.1038/ng.3721.
58. Creyghton MP, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 2010;107:21931–6. https://doi.org/10.1073/pnas.1016071107.
59. Hess J, Angel P, Schorpp-Kistner M. AP-1 subunits: quarrel and harmony among siblings. J Cell Sci. 2004;117:5965–73. https://doi.org/10.1242/jcs.01589.
60. Gaunt TR, et al. Systematic identification of genetic influences on methylation across the human life course. Genome Biol. 2016;17:61. https://doi.org/10.1186/s13059-016-0926-z.
61. Richmond RC, Suderman M, Langdon R, Relton CL, Davey Smith G. DNA methylation as a marker for prenatal smoke exposure in adults. Int J Epidemiol. 2018;47:1120–30. https://doi.org/10.1093/ije/dyy091.
62. Tajuddin SM, et al. Novel age-associated DNA methylation changes and epigenetic age acceleration in middle-aged African Americans and whites. Clin Epigenet. 2019;11:119. https://doi.org/10.1186/s13148-019-0722-1.
63. Horvath S, et al. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. Genome Biol. 2016;17:171. https://doi.org/10.1186/s13059-016-1030-0.
64. Liu Z, et al. The role of epigenetic aging in education and racial/ethnic mortality disparities among older U.S. Women. Psychoneuroendocrinology. 2019;104:18–24. https://doi.org/10.1016/j.psyneuen.2019.01.028.
65. Philibert R, et al. Array-based epigenetic aging indices may be racially biased. Genes (Basel). 2020. https://doi.org/10.3390/genes11060685.
66. Wang Y, et al. Preoperative neutrophil-to-lymphocyte ratio predicts response to first-line platinum-based chemotherapy and prognosis in serous ovarian cancer. Cancer Chemother Pharmacol. 2015;75:255–62. https://doi.org/10.1007/s00280-014-2622-6.
67. Ozcan C, et al. The prognostic significance of preoperative leukocytosis and neutrophil-to-lymphocyte ratio in patients who underwent radical cystectomy for bladder cancer. Can Urol Assoc J. 2015;9:E789-794. https://doi.org/10.5489/cuaj.3061.
68. Salim DK, et al. Neutrophil to lymphocyte ratio is an independent prognostic factor in patients with recurrent or metastatic head and neck squamous cell cancer. Mol Clin Oncol. 2015;3:839–42. https://doi.org/10.3892/mco.2015.557.
69. Ambatipudi S, et al. Assessing the role of DNA methylation-derived neutrophil-to-lymphocyte ratio in rheumatoid arthritis. J Immunol Res. 2018;2018:2624981. https://doi.org/10.1155/2018/2624981.
70. Cronje HT, et al. Methylation vs. protein inflammatory biomarkers and their associations with cardiovascular function. Front Immunol. 2020;11:1577. https://doi.org/10.3389/fimmu.2020.01577.
71. Azab B, Camacho-Rivera M, Taioli E. Average values and racial differences of neutrophil lymphocyte ratio among a nationally representative sample of United States subjects. PLoS ONE. 2014;9: e112361. https://doi.org/10.1371/journal.pone.0112361.
72. Bergstedt J, et al. Factors driving DNA methylation variation in human blood. bioRxiv. 2021. https://doi.org/10.1101/2021.06.23.449602.
73. Munafo MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. Int J Epidemiol. 2018;47:226–35. https://doi.org/10.1093/ije/dyx206.
74. Bird PK, et al. Growing up in Bradford: protocol for the age 7–11 follow up of the Born in Bradford birth cohort. BMC Public Health. 2019;19:939. https://doi.org/10.1186/s12889-019-7222-2.
75. Min JL, Hemani G, Davey Smith G, Relton C, Suderman M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. Bioinformatics. 2018;34:3983–9. https://doi.org/10.1093/bioinformatics/bty476.
76. Reinius LE, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS ONE. 2012;7:e41361. https://doi.org/10.1371/journal.pone.0041361.

77. Das S, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48:1284–7. https://doi.org/10.1038/ng.3656.
78. Genomes Project, C., et al. A global reference for human genetic variation. Nature. 2015;526:68–74. https://doi.org/10.1038/nature15393.
79. Chang CC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7. https://doi.org/10.1186/s13742-015-0047-8.
80. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115. https://doi.org/10.1186/gb-2013-14-10-r115.
81. Hannum G, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49:359–67. https://doi.org/10.1016/j.molcel.2012.10.016.
82. Alfonso G, Gonzalez JR. Bayesian neural networks for the optimisation of biological clocks in humans. BioRxiv. 2020. https://doi.org/10.1101/2020.04.21.052605.
83. Horvath S, et al. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. Aging (Albany NY). 2018;10:1758–75. https://doi.org/10.18632/aging.101508.
84. Levine ME, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY). 2018;10:573–91. https://doi.org/10.18632/aging.101414.
85. Lu AT, et al. DNA methylation-based estimator of telomere length. Aging (Albany NY). 2019;11:5895–923. https://doi.org/10.18632/aging.102173.
86. Suderman M, et al. Dmrff: identifying differentially methylated regions efficiently with power and control. BioRxiv. 2018. https://doi.org/10.1101/508556.
87. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164. https://doi.org/10.1093/nar/gkq603.
88. Johnson KC, Houseman EA, King JE, Christensen BC. Normal breast tissue DNA methylation differences at regulatory elements are associated with the cancer risk factor age. Breast Cancer Res. 2017;19:81. https://doi.org/10.1186/s13058-017-0873-y.
89. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9. https://doi.org/10.1038/75556.
90. The Gene Ontology, C. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 2017;45:D331–8. https://doi.org/10.1093/nar/gkw1108.
91. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics. 2016;32:286–8. https://doi.org/10.1093/bioinformatics/btv560.
92. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics. 2016;32:587–9. https://doi.org/10.1093/bioinformatics/btv612.

Publisher's Note